Applied Animal Behaviour Science 163 (2015) 158–166

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

Consistency of shelter dogs' behavior toward a fake versus real stimulus dog during a behavior evaluation

Anastasia Shabelansky[a,*], Seana Dowling-Guyer[a,b], Hilary Quist[b], Sheila Segurson D'Arpino[a], Emily McCobb[b]

[a] Center for Shelter Dogs, Animal Rescue League of Boston, 10 Chandler Street, Boston, MA 02116, USA
[b] Center for Animals and Public Policy, Cummings School of Veterinary Medicine, Tufts University, North Grafton, MA 01536, USA

ARTICLE INFO

Article history:
Accepted 5 December 2014
Available online 13 December 2014

Keywords:
Fake dog
Shelter dogs
Behavior evaluation
Temperament test
Dog to dog aggression

ABSTRACT

Behavior evaluations are widely used by animal shelters and other organizations that rehome dogs. The dog-to-dog subtest is a common feature of most canine behavior evaluations. The use of model devices such as a stuffed dog during this subtest could be convenient for shelters and increase safety. However, there is little research indicating if a fake dog can be reliably used instead of a live dog. In this study, the consistency of shelter dogs' reactions toward a fake and a real dog during the dog-to-dog subtest was investigated. Forty-five shelter dogs were evaluated using two different stimulus conditions. In one condition, the test dog was confronted with a single plush dog (the same plush dog for all test dogs), and in the other, with a single live dog (the same live dog for all test dogs). A standardized list of behaviors was recorded as observed or absent for both conditions with each dog serving as its own control. To calculate the agreement of individual behaviors between the two conditions, Cohen's Kappa was used. However, since many of the behaviors occurred at very low or high frequency rates, Prevalence-Adjusted, Bias-Adjusted Kappa (PABAK) was used along with Cohen's Kappa due to Cohen's Kappa's sensitivity to high or low prevalence, for which PABAK adjusts. For the purposes of this study, PABAK or Kappa scores greater than 0.61 were considered an indicator of a good degree of agreement between reactions toward the fake and the real dogs. The degree of agreement varied widely across individual behaviors with, Kappa ranging from −0.04 to 0.75 and PABAK from 0.29 to 1. Collapsing individual behaviors into behavior traits (e.g., friendly, aggressive, fearful) revealed a high degree of agreement for the friendly trait (Kappa = 0.60, PABAK = 0.69). However, the aggressive trait did not demonstrate adequate agreement (Kappa = 0.11 and PABAK = 0.38) and the fearful trait demonstrated only moderate agreement between the two stimulus conditions (Kappa = 0.50 and PABAK = 0.51). These results suggest that, while it may be possible to use a fake dog for the dog-to-dog subtest to assess friendly behavior toward other dogs, fearful and aggressive behaviors may not be consistent between the fake and real dogs, thus limiting the usefulness of the fake dog during behavior evaluations. In addition, the results of this study suggest more research is needed into the predictive validity of both fake and real dogs, since it appears the stimulus dog, whether fake or real, can influence the subtest's results.

* Corresponding author. Tel.: +1 617 226 5604; fax: +1 617 226 5679.
E-mail address: ashabelansky@gmail.com (A. Shabelansky).

http://dx.doi.org/10.1016/j.applanim.2014.12.001
0168-1591/

1. Introduction

Behavior evaluations play a critical role for shelter and rescue dogs, and are used to identify behavior tendencies in order to rehome an animal into an appropriate home (D'Arpino et al., 2012; Dowling-Guyer et al., 2011; Ledger and Baxter, 1997; Reid and Collins, 2012; Van der Borg et al., 1991). Research has shown that 28% of sheltering organizations use a standardized behavior evaluation (D'Arpino et al., 2012), with 63% of higher volume organizations (with annual intake more than 1000 dogs) using one. According to Taylor and Mills (2006), the dog-to-dog subtest is one of the 20 most commonly used subtests included in most standardized behavior evaluations, such as: Assess-A-Pet™ (Bollen and Horowitz, 2008), Match-Up II Shelter Dog Rehoming Program™ (Center for Shelter Dogs, 2013; Dowling-Guyer et al., 2011), SAFER® Aggression Assessment (Weiss, 2007). This subtest is used to gain insight into shelter dog behavior toward conspecifics. Since many shelter dogs are strays or transfers from external organizations, with no behavioral history to consult, the dog-to-dog subtest is a key instrument to gather information about a dog's behavior toward other dogs. Although common, it can be a difficult subtest to implement because it depends entirely on what other dogs are available in the shelter for use as the stimulus dog. Therefore, the use of model devices such as a stuffed dog could be beneficial for many shelters. Model devices are convenient to use and reduce the risk of injury to the dog, and might be especially useful in large organizations where staff has limited time to perform evaluations. Moreover, use of a fake dog instead of a real stimulus dog during a dog-to-dog evaluation allows direct interaction that otherwise could be uncomfortable or dangerous for both the dog and handler (Reid and Collins, 2012).

Although the use of a life-like artificial dog substitute in the evaluation of shelter dogs is not a new idea, there is little research indicating if a fake dog can be reliably used instead of a live dog in shelter behavior evaluations. Previous studies have used plastic dogs (Barnard et al., 2012; Reid and Collins, 2012) or stuffed models (Leaver and Reimchen, 2008) to learn how dogs behave toward these models. In a study by Leaver and Reimchen (2008), a fake dog was fitted with several varying-length tails and introduced to several real dogs to learn more about canine communication. Two other studies evaluated the validity of using a fake dog during a behavior evaluation. Barnard et al. (2012) used the C-BARQ questionnaire to assess the dogs' behavior history toward conspecifics, then the same dogs were evaluated using a plastic dog. Although the fake dog elicited many social reactions that were not observed when an ambiguous object (black plastic garbage bag filled with crumpled newspaper) was shown to the dogs, the reactions to the device were only partially consistent with the dogs' aggressive history. Only two dogs from the group of 12 dogs who had a history of aggression toward conspecifics displayed some levels of aggressive behavior (bark, growl, snaps) toward the fake dog. Reid and Collins (2012) evaluated Pit Bull Terrier type fighting dogs. They found that the dogs' responses to the fake dog were similar to a real dog using the same evaluation. However, incidences of aggression toward conspecifics in dogs bred for fighting are significantly higher than in most shelter dogs (Capra et al., 2009; Reid and Collins, 2012). Consequently, Reid and Collins's study results might not apply to a shelter population which is more diverse than the dogs used in that study.

This study investigated the subject further using shelter dogs, a controlled environment and a standardized dog-to-dog interaction assessment to examine the consistency of reactions toward a fake and a real dog. We hypothesized that dogs would react similarly to a fake dog as to a real dog. This study was not designed to investigate whether information gathered during the dog-to-dog subtest accurately predicted future behavior toward conspecifics; rather, this study examined the consistency of behavior between the two conditions. The results could advise shelters whether they could reliably use a fake dog instead of a live dog in dog-to-dog subtests already being conducted in shelters. In addition, information about the effectiveness of the fake dog as a proxy for live dogs may help shelters to accelerate the evaluation process, as well as make it safer and more comfortable for both humans and animals.

2. Materials and methods

The behavior evaluation and the subsequent analysis of test results received full ethical approval from participating shelter organizations' senior management. The protocol for this study was approved by the Center for Shelter Dogs' internal review panel prior to initiation of the project.

2.1. Subjects

The sample for this study consisted of 45 shelter dogs from two shelters in Massachusetts, USA. To be eligible for inclusion, dogs had to be 6 or more months of age on the first day of the study, "apparently" healthy, without medical conditions which would restrict their physical activity, able to be handled by trained research staff and in the shelter for at least 72 h before the study began.

Three shelter dogs who were included in the study were in foster care with shelter staff. These dogs were office fostered during the day and fostered at home during the night. Since the dogs were technically in the shelter for 72 h and since office fostering is an acceptable routine in many shelters, these dogs were included in the study. Of the 45 dogs included in the study, 6 were intact females, 13 were spayed females, 6 were intact males and 17 were neutered males. Three female dogs had an unconfirmed spayed status at the time of the study. It is unknown how many dogs were actually purebred as neither pedigrees nor DNA analyses were available. The predominant breed was assigned by the shelter staff based upon phenotype. The most commonly assigned predominant breeds were: Pit Bull type (17), Labrador Retriever (6), German Shepherd (5), Dachshund (3), Boxer (2), Chihuahua (2), Lhasa Apso (2), Siberian Husky (2). Median weight range was 18–22 kg, and mean (SD) age was 2.4 years (1.7 years).

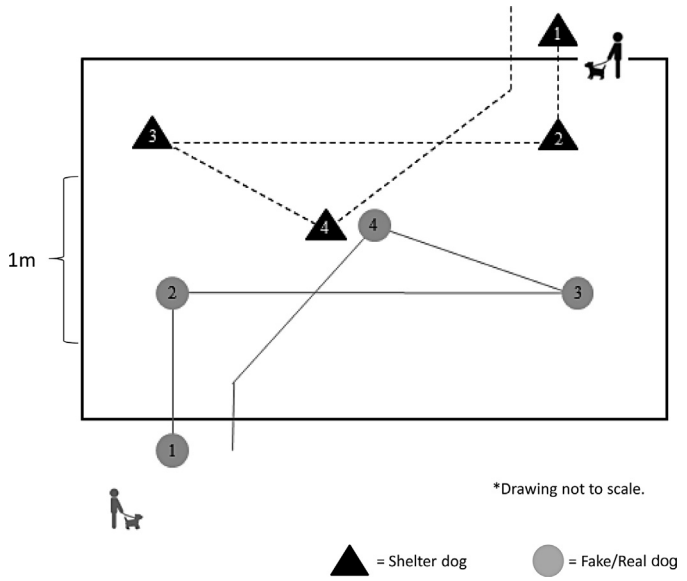

Fig. 1. Schematic view of movement steps performed by two handlers during the assessment; see text for further explanation.

2.2. Study sites

Shelter dogs from two private animal shelters, one urban and one rural, were evaluated. The rural testing area was an outdoor dirt area approximately 7 m × 10 m enclosed on three sides by a 2 m high chain-link fence with the fourth wall being an external wall of the shelter. Test dogs entered the outdoor fenced area through a chain-link gate approximately 1 m wide that was external to the shelter. At the urban animal shelter, evaluations took place in a large auditorium approximately 6.1 m x 9.1 m with linoleum tile flooring. Test dogs entered through a pair of double doors at the corner of the room external to the shelter.

2.3. Dog-to-dog assessment

Twenty-five dogs were evaluated in the yard at the rural location and 20 dogs were evaluated in the auditorium at the urban location. Prior to entering the evaluation location, dogs were walked on a leash outside allowing time to urinate and defecate, using praise and treats to reward calm behavior while walking. The same scenario was repeated twice for each dog using two different stimulus conditions. In one condition, the dog was confronted with a fake, plush dog and in the other with a real dog. The real stimulus dog was a neutered male, American Staffordshire terrier mix, approximately 5 years in age, dark brown/tan brindle coat, with a demonstrated history of friendly interactions with other dogs. The fake stimulus dog had the appearance of a pointer breed dog, with a white body and several black and brown spots on its back and face, posed in a standing position with head facing forward. The fake dog's dimensions were 35.5 cm × 58.4 cm × 76.2 cm, and weighed 2 kg. This particular fake dog was chosen based upon its similarity in size to the real dog. The presentation of the fake and real dogs was alternated every other time, so half of the sample was introduced to the fake dog first and half of the sample to the real dog first. In addition, the presentation order to the test dogs was balanced based on sex, to make sure that the same amount of males and females were introduced to the fake dog first and vice versa. There was a 5 min break between scenarios, during which the test dog was allowed to run unleashed in the evaluation area. All dogs, including the fake and the real dog, wore a martingale collar attached to a 2 m leash. The real stimulus dog was monitored closely for any signs of stress or exhaustion, at which point testing stopped and continued on a different day. The testing was stopped three times due to stimulus dog exhaustion, resulting in a total of 3 testing days.

The dog-to-dog assessment was taken from the larger Match-Up II Behavior Evaluation, a standardized test which measures a dog's behavioral reactions to a series of scenarios, or subtests. Match-Up II evaluates dog behavioral tendencies and personality by scoring observed (marked 1) or not observed (marked 0) individual behaviors, body postures and movement (Center for Shelter Dogs, 2013; Dowling-Guyer et al., 2011; Marder et al., 2013). Two people carried out the assessment, one person holding the real and then the fake dog (handler 1) and the other one holding the test dog (handler 2). In addition, there was a videographer, a rater and an assistant in the room who helped with dogs' rotation and distributed scoring sheets to the rater. Although the video recordings were collected, they were not used in the study. The procedure was as follows: (1) handler 2 walked the test dog to the evaluation area where the real or the fake dog was already present in the opposite corner with handler 1; (2) handlers allowed both the test and the real/fake dog to pass each other approximately 1 m apart on the leash, without touching; (3) dogs were allowed to once again walk toward each other and meet in the center of the evaluation area for 20 s, direct contact was allowed, including

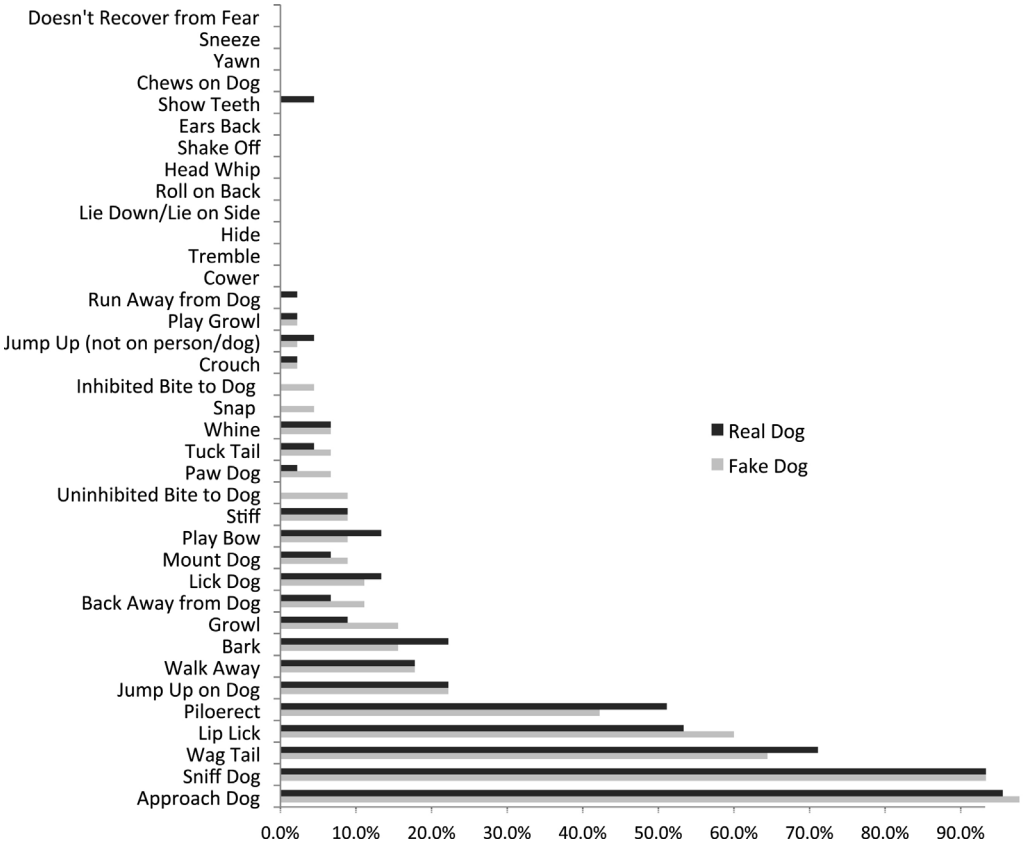

Fig. 2. Percent of dogs displaying different reactions toward fake and real stimulus dogs.

any natural circling from head to rear, on a leash as loose as possible; (4) the test dog was guided apart by the handler, with the real/fake dog exiting the evaluation area. The fake dog was made to move in an animated manner with the handler holding the leash with one hand and the tail with the other while floating, and not bouncing, the dog along the ground. The fake dog was made to circle and slightly approach the test dog during the circling portion of the assessment. The duration of each testing scenario was about 1 min. If any signs of aggression occurred during the second or third part, the assessment was stopped and the dogs separated. If aggression was present during the first part, the dog was still walked by the real/fake dog 1 m apart; if aggressive behavior persisted, the assessment was stopped and the real/fake dog left the room. The second scenario for the same dog was performed in the same manner regardless of aggressive behavior on the first scenario. For those dogs who displayed aggressive behaviors, reactions were displayed at the end of the assessment, consequently, none of the assessments were stopped, resulting in 1 min average time for each scenario for every dog. The scheme of both dogs' movements is presented in Fig. 1.

The rater recorded 37 specific behaviors during the actual assessment (Fig. 2). Behaviors were scored as observed or not observed on a standardized scoring form. Each behavior was independently scored as present or absent, regardless of what other behaviors were observed, meaning the dog could 'approach' the fake/real dog initially but then 'walk away'. The rater was a Certified Pet Dog Trainer (CPDT) with experience with dog evaluations. Both rater and handler were blinded to the test dogs' behavior history. The rater went through a short training in which the study design, scoring sheet, and definitions of behaviors were discussed.

2.4. Data analysis

The percentage of dogs exhibiting each behavior toward the fake and the real dogs was calculated. The percentage of dogs exhibiting consistent behaviors (percent agreement) between the fake and the real dog was calculated for each behavior. Positive and negative agreement as well as Cohen's Kappa (Cohen, 1960) were generated. Positive and negative agreement provide information about the degree of agreement separately for positive and negative occurrences, which often is a more informative indication of agreement than the overall percentage agreement (Uebersax, 2013). In fact, Kappa is a weighted sum of these two quantities. Different from Kappa, these indices can reveal agreement on the absence of behavior, but not on its presence and vice versa. For example, approach was a very prevalent behavior with 93% of dogs approaching both the fake and real dogs, leading to a high positive agreement

Table 1
Individual behaviors grouped into behavior traits.

Aggressive	Fearful	Friendly/playful
Growl	Back away from dog	Jump up on dog
Inhibited bite	Crouch	Lick dog
Show teeth	Lip lick	Paw dog
Snap	Run away from dog	Play bow
Stiff	Tuck tail	Play growl
Uninhibited bite		Wag tail

Only observed behaviors included.

of 97%. However, 7% of dogs approached only either the fake or real dog: one approached the fake dog only and two approached the real dog only. Consequently, there were zero dogs who did not approach both the fake and real dogs, leading to 0% negative agreement. Another example is "stiff". Eighty-four percent of dogs (38) did not stiffen when they saw both the fake and real dogs, leading to a high negative agreement between the two conditions (93%). Three dogs stiffened when they saw the fake dog but not the real dog and three dogs stiffened when they saw the real dog but not the fake dog. Only one dog (2%) of 45 stiffened when it saw both the fake and real dogs, leading to a very low positive agreement (25%). Since some behaviors were rarely or very frequently observed, the Prevalence-Adjusted, Bias-Adjusted Kappa (PABAK) coefficient was calculated for each behavior variable (Mackinnon, 2000), in addition to Cohen's Kappa, since bias and prevalence affects Cohen's Kappa. The guidelines for interpretation of the kappa coefficients suggested by Landis and Koch (1977) are as follows: Kappas $\leqslant 0$ represent poor agreement, from 0.01 to 0.20 slight agreement, from 0.21 to 0.40 fair agreement, from 0.41 to 0.60 moderate agreement, from 0.61 to 0.80 substantial agreement, and from 0.81 to 1.00 almost perfect agreement. For the purposes of this study, PABAK or Kappa scores greater than 0.61 were considered an indicator of a good degree of agreement between reactions toward the fake and the real dogs. When interpreting the results, the following guidelines were used: for behaviors which were not rare or extremely common (approximately 20–80% prevalence), the positive and negative agreement, as well as Cohen's Kappa and PABAK were examined. When prevalence rates were very high or low, PABAK as well as positive and negative agreement provided more useful information (Uebersax, 2013).

After review of the individual behaviors, the majority of behaviors were grouped together to create three behavior traits: fearful, aggressive, and friendly/playful (Table 1), based on previous research (De Palma et al., 2005; Dowling-Guyer et al., 2011; Svartberg, 2007). Grouping behaviors into traits can provide more insight into the overall agreement of a dog's reactions between the two conditions. For example, if the dog bit the fake dog but showed less severe reactions such as growl toward the real dog, grouping growl and bite under the same aggressive trait provides a common ground for comparison of aggressive behavior between the two stimulus conditions. A data reduction technique was not used in this study because previous work has already been done with larger samples which statistically demonstrated which behaviors load together into conceptually related traits (Dowling-Guyer et al., 2011). Dogs were classified as showing one of the traits if they exhibited at least one behavior from the trait. Since each behavior was independently scored, traits were not mutually exclusive or antithetical. Thus, a dog who showed 'bite' and 'wag tail' could be classified under both aggressive and friendly traits. Consequently, a dog could be categorized under any one or more than one trait. Dogs who did not show any of the grouping behaviors were categorized as having none of these traits. Behaviors with zero prevalence were excluded from grouping. 'Approach' and 'sniff' were excluded due to their very high prevalence as the design of the assessment resulted in most dogs demonstrating the behaviors 'approach' and 'sniff'. The consistency of behavior traits was analyzed between the two dog conditions, using the same indices as described above.

Descriptive statistics were generated and data were analyzed using IBM SPSS Statistics, version 20. Dag_Stat was used to calculate agreement percentages, Cohen's Kappa, and PABAK (Mackinnon, 2000). Pearson's Chi-square test was used to identify whether the order of the fake dog presentation was significantly associated with presence and absence of aggressive behavior (Siegel, 1956). The 2×2 table was used with "order" variable on one side (fake dog presented first/real dog presented first) and "aggression" (yes/no) variable on another side.

2.5. Reliability analysis

Interrater reliability was not analyzed in this study because previous work had already been done with a larger sample using the same dog-to-dog assessment described above. Interrater reliability for the dog-to-dog assessment was calculated as a part of this previous work, the goal of which was to investigate the interrater reliability of individual behavior scoring. The dog-to-dog assessment for the larger project was performed in the same manner as described in this study. Two experienced dog handlers observed 50 live dog-to-dog assessments of shelter dogs and coded for the presence or absence of 30 behaviors. Of the 30 behaviors, eight (26.7%) achieved perfect agreement. For the remaining 22 behaviors, PABAK scores ranged from 0.52 to 0.96, with a mean of 0.84. Since many of the behaviors occurred at very low or high frequency rates, PABAK was used in place of Cohen's Kappa due to Cohen's Kappa's sensitivity to high or low prevalence, for which PABAK adjusts (Mackinnon, 2000). With the exception of the behavior "lip lick" which achieved the lowest PABAK score of 0.52, all other behaviors achieved a minimum of 0.60 or higher degree of agreement.

3. Results

3.1. Behavior frequencies

Fig. 2 presents the percent of dogs who exhibited each behavior toward the fake and the real dog separately. A wide range of behaviors was observed, although 'approach'

Table 2
Agreement on individual behaviors of the test dogs between presentation of live and fake dogs during dog-to-dog evaluation.

Behavior	% Observed agreement	Positive agreement	Negative agreement	Cohen's Kappa	PABAK
Approach dog	93%	0.97 (CI: 0.93–1.00)	0.00	−0.03 (CI: −0.07–0.01)	0.87
Back away from dog	91%	0.50 (CI: 0.08–0.92)	0.95 (CI: 0.90–1.00)	0.46 (CI: 0.01–0.90)	0.82
Bark	84%	0.59 (CI: 0.31–0.87)	0.90 (CI: 0.83–0.97)	0.50 (CI: 0.18–0.81)	0.69
Chews on dog	100%	–	1.00	–	–
Cower	100%	–	1.00	–	–
Crouch	96%	0.00	0.98 (CI: 0.95–1.01)	−0.02 (CI: −0.05–0.01)	0.91
Doesn't recover from fear within 30 s	100%	–	1.00	–	–
Ears back	100%	–	1.00	–	–
Growl	80%	0.18 (CI: −0.13–0.49)	0.89 (CI: 0.81–0.96)	0.08 (CI: −0.25–0.41)	0.60
Head whip	100%	–	1.00	–	–
Hide	100%	–	1.00	–	–
Inhibited bite to dog	96%	0.00	0.98 (CI: 0.95–1.01)	–	0.91
Jump up (not on person/dog)	93%	0.00	0.97 (CI: 0.93–1.00)	−0.04 (CI: −0.07–0.01)	0.87
Jump up on dog	78%	0.50 (CI: 0.23–0.77)	0.86 (CI: 0.77–0.94)	0.36 (CI: 0.04–0.68)	0.56
Lick dog	84%	0.36 (CI: 0.00–0.73)	0.91 (CI: 0.85–0.98)	0.28 (CI: −0.12–0.67)	0.69
Lie down/lie on side	100%	–	1.00	–	–
Lip lick	76%	0.78 (CI: 0.66–0.91)	0.72 (CI: 0.56–0.88)	0.50 (CI: 0.25–0.76)	0.51
Mount	93%	0.57 (CI: 0.13–1.01)	0.96 (CI: 0.92–1.00)	0.54 (CI: 0.07–1.00)	0.87
Paw dog	91%	0.00	0.95 (CI: 0.91–1.00)	−0.03 (CI: −0.09–0.02)	0.82
Piloerect	64%	0.62 (CI: 0.45–0.79)	0.67 (CI: 0.51–0.82)	0.29 (CI: 0.02–0.57)	0.29
Play growl	100%	1.00	1.00	1.00	1.00
Playbow	91%	0.60 (CI: 0.24–0.96)	0.95 (CI: 0.90–1.00)	0.55 (CI: 0.17–0.94)	0.82
Roll on back	100%	–	1.00	–	–
Run away from dog	98%	0.00	0.99 (CI: 0.97–1.00)	–	0.96
Shake off	100%	–	1.00	–	–
Show teeth	96%	0.00	0.98 (CI: 0.95–1.01)	–	0.91
Snap	96%	0.00	0.98 (CI: 0.95–1.01)	–	0.91
Sneeze	100%	–	1.00	–	–
Sniff dog	91%	0.95 (CI: 0.91–1.00)	0.33 (CI: −0.15–0.82)	0.29 (CI: −022–0.79)	0.82
Stiff	87%	0.25 (CI: −0.15–0.65)	0.93 (CI:0.87–0.99)	0.18 (CI: −0.24–0.60)	0.84
Tremble	100%	–	1.00	–	–
Tuck tail	93%	0.40 (CI: −0.14–0.94)	0.96 (CI: 0.92–1.00)	0.37 (CI:−0.19–0.93)	0.87
Uninhibited bite to dog	91%	0.00	0.95 (CI: 0.91–1.00)	–	0.82
Wag tail	89%	0.92 (CI: 0.68–0.98)	0.83 (CI: 0.85–0.99)	0.75 (CI: 0.54–0.95)	0.78
Walk away	82%	0.50 (CI: 0.20–0.80)	0.89 (CI: 0.82–0.97)	0.39 (CI: 0.05–0.74)	0.64
Whine	96%	0.67 (CI: 0.23–1.10)	0.98 (CI: 0.94–1.01)	0.64 (CI: 0.18–1.10)	0.91
Yawn	100%	–	1.00	–	–

Confidence intervals (CI) were calculated at the 95% confidence level.
Observed agreement is an overall agreement between two raters on the presence and absence of behavior, with 0% meaning no agreement at all and 100% meaning perfect agreement.
Positive agreement is an agreement between raters on the presence of behavior with 0 meaning no agreement at all and 1 meaning perfect agreement, where observers agree every time the behavior occurs.
Negative agreement is an agreement between raters on the absence of behavior with 0 meaning no agreement at all and 1 meaning perfect agreement, where observers agree every time the behavior does not occur.
Kappa higher than 0.61 is an indicator of a good degree of agreement between reactions toward the fake and the real dogs.
PABAK higher than 0.61 is an indicator of a good degree of agreement between reactions toward the fake and the real dogs.

and 'sniff dog' were the most common behaviors, displayed by more than 90% of the sample toward both the fake and the real dog. The second most common behaviors were 'wag tail' and 'lip lick', displayed by more than 50% of the sample, toward both the fake and the real dog. However, some behaviors were not observed at all, among them: 'cower', 'tremble', 'hide', 'lie down/lie on side', 'roll on back', 'head whip', 'shake off', 'ears back', 'chews on dog', 'yawn', 'sneeze', 'doesn't recover from fear'.

3.2. Agreement on individual behaviors between two conditions

The percent of observed agreement between the two stimulus scenarios was very high for all behaviors, ranging from 64% to 100% (Table 2). However, for many behaviors, the percent of agreement was impacted by very low or very high prevalence. For example, 'growl' was exhibited by only 8.9% (four dogs) of dogs toward the real dog and by 15.6% (seven dogs) of dogs toward the fake dog but has a consistent result of 80% between the two conditions. However, the low positive agreement of 0.18 reveals that there was very little agreement between two stimuli conditions on the presence of 'growl'. Meaning, if a dog growled toward the fake dog, it did not necessarily growl toward the real dog and vice versa. In addition, 'growl' shows a high degree of negative agreement (0.89), meaning if 'growl' was not observed toward the fake dog, it most likely was not observed toward the real dog. Very similarly, behaviors that were not observed at all resulted in an observed

Table 3
Agreement on grouped behvaior traits of the test dogs between presentation of live and fake dogs during dog-to-dog evaluation.

Behavior traits	% Observed agreement	Positive agreement	Negative agreement	Cohen's Kappa	PABAK
Aggressive	69%	0.30 (CI: 0.04–0.56)	0.80 (CI: 0.70–0.90)	0.11 (CI: −0.19–0.41)	0.38
Fearful	76%	0.79 (CI: 0.60–0.87)	0.70 (CI: 0.53–0.87)	0.50 (CI: 0.25–0.75)	0.51
Friendly/playful	84%	0.90 (CI: 0.82–0.97)	0.70 (CI: 0.48–0.91)	0.60 (CI: 0.34–086)	0.69

Confidence intervals (CI) were calculated at the 95% confidence level.
Observed agreement is an overall agreement between two raters on the presence and absence of behavior, with 0% meaning no agreement at all and 100% meaning perfect agreement.
Positive agreement is an agreement between raters on the presence of behavior with 0 meaning no agreement at all and 1 meaning perfect agreement, where observers agree every time the behavior occurs.
Negative agreement is an agreement between raters on the absence of behavior with 0 meaning no agreement at all and 1 meaning perfect agreement, where observers agree every time the behavior does not occur.
Kappa higher than 0.61 is an indicator of a good degree of agreement between reactions toward the fake and the real dogs.
PABAK higher than 0.61 is an indicator of a good degree of agreement between reactions toward the fake and the real dogs.

agreement of 100% which was solely based on negative agreement. Among behaviors that had prevalence rates ranging between 20% and 80%, only 'wag tail' had a substantial degree of agreement (Kappa = 0.75, PABAK = 0.78). 'Lip lick' showed moderate agreement between two conditions (Kappa = 0.50, PABAK = 0.52). 'Jump up on dog' showed a high degree of negative agreement (0.86) but a lower degree of positive agreement (0.50) with Kappa = 0.36 and PABAK = 0.56. Piloerect demonstrated only fair agreement between the two stimulus conditions with Kappa = 0.29 and PABAK = 0.29, and both positive (0.62) and negative (0.67) agreements being relatively low. 'Bark' had a prevalence rate of 22% when the real dog was introduced and 16% when the fake dog was introduced, and had a moderate degree of agreement (Kappa = 0.50, PABAK = 0.69, positive agreement = 0.59, negative agreement = 0.90).

3.3. Agreement on behavior traits between two conditions

After grouping into behavior traits (Table 3), 17 dogs (38%) showed some form of aggressive behavior toward either the real or fake dog. Among them, eight dogs showed aggressive behavior toward the real dog, with four of them exhibiting only 'stiff' and the rest displaying 'growl'. Twelve dogs exhibited aggression toward the fake dog, with six (50%) of them biting it. Only three dogs were aggressive toward both the fake and real dog. Consequently, adequate agreement was not found between aggressive behavior toward the fake and real dog (Kappa = 0.11, PABAK = 0.38, positive agreement = 0.30, negative agreement = 0.80). Although observed agreement between the two stimulus conditions was 69% (31/45), it was largely dominated by negative agreement, when aggression was not observed in either stimulus conditions. There was no presentation order effect on aggressive behavior ($X^2 = 0.008$, $P = 0.93$). When the fake dog was shown first, six dogs showed aggressive behaviors (any combination of growl, inhibited bite, show teeth, snap, stiff, uninhibited bite). When the real dog was shown first, six dogs showed a combination of the above listed aggressive behaviors.

Thirty-two (71%) of 45 dogs showed fear in one or both stimulus conditions. Twenty-one (66%) of them showed fear toward both the fake and real dogs, eight (25%) dogs showed fear toward the fake dog only and only three (9%) dogs showed fear toward the live dog only. The fearful trait demonstrated only moderate agreement between the two stimulus conditions with an observed agreement of 76% (34/45). Positive agreement was 0.79 and negative agreement was 0.70, with Kappa = 0.50 and PABAK = 0.51.

The majority of dogs (37/45) showed friendly behavior in one or both stimulus conditions with 67% (30/37) of them exhibiting friendly behavior toward both the fake and real dogs. Positive agreement between two stimulus conditions was 0.90 and negative agreement was 0.70. Kappa showed a moderate agreement (Kappa = 0.60) but PABAK revealed substantial agreement (PABAK = 0.69) since it corrects for the relatively high prevalence of friendly behavior.

4. Discussion

Our hypothesis that dogs would react similarly to a fake dog as to a real dog was partially supported by the results of this study. Although the observed agreement was very high for all behaviors, it was mainly due to the absence of behaviors rather than to their presence. Because so few instances of many behaviors were observed, those behaviors reached a very high degree of negative agreement, meaning that when the behavior was not present in one condition, it most likely was not present in the other condition. Conversely, 'approach' and 'sniff dog' reached a high degree of positive agreement but both were very prevalent, exhibited by more than 93% percent of the sample. However, when examining behaviors with prevalence rates between 20% and 80%, only 'wag tail' reached an adequate degree of agreement, suggesting that there are limitations to the consistency of reaction to a fake and real dog.

We found that the fake dog elicited more aggressive reactions than the real dog although the difference was not statistically significant. One possible explanation is that the fake dog was walked directly toward the test dog with its head up, looking forward, which may have elicited aggressive behaviors, as opposed to the real dog which may have avoided any aggressive signaling toward the test dog. The tendency to display more aggressive behavior toward the fake dog was also noted by Collins et al. (2012). In that study, the authors observed pit bull puppies displaying more shaking, holding, and biting behaviors toward the fake puppy than toward littermates or playmates. However, they attributed it to the puppies perceiving the

fake device as a toy. Another explanation for the greater frequency and severity of aggressive behavior displayed toward the fake dog could be that the stimulus dog's behavior acted to decrease an aggressive response in the test dog. This idea is supported by previous research showing that it is possible for stimulus dog behavior to influence an aggressive response in tested dogs (Gazzano et al., 2010; Mariti et al., 2010; Netto and Planta, 1997). Netto and Planta (1997) developed a dog evolution test in which they presented various stimuli that are known to elicit aggression in dogs, including other dogs. To perform a predictability analysis of the test, they re-tested 37 dogs from the original sample. They found that the subtests in which dogs were used as stimuli had the lowest predictability. Since the study staff prevented the stimulus dogs from being "defeated" by the tested dogs, male stimulus dogs reacted more aggressively and the female stimulus dog behaved less aggressively, as the test continued. The authors suggest that inconsistent behavior on the part of the stimulus dogs could have had an impact on tested dogs, resulting in lower predictability rates of dog-to-dog subtest. Moreover, for future studies, the authors suggest monitoring stimulus dogs closely and replacing them with another stimulus dog if they become too aggressive. In the current study, the real stimulus dog was carefully chosen to be extremely friendly; during the test, it was noted that the stimulus dog appeared to be responding to some aggressive signals from the test dogs by turning his head aside and avoiding looking at the test dog. "Calming signals" such as described above have been shown to have a communicative and calming effect on receiving dogs. In Gazzano et al. (2010), a 5 min meeting between 20 dogs was performed. Every dog met four other dogs differing in gender and familiarity. Researchers examined 21 different signals which occurred 1213 times during the interactions. "Turing Head" and "Looking Elsewhere" were the most commonly displayed signals. They observed that on 62 occasions, test dogs showed an aggressive behavior before they received calming signals from stimulus dogs and, in 77% of cases, these signals reduced the test dogs' aggression. In a different study (Mariti et al., 2010), it was found that a higher number of "calming signals", in particular "Looking Elsewhere" and "Freezing", were displayed when the test dog met an unfamiliar dog. The authors suggest that "calming signals" play a critical role in reducing the chance of aggressive displays. Since the live stimulus dog in our study could alter his behavior and calm the testing dog whereas the fake dog could not, it is possible that the stimulus dog actually influenced the test dog's behavior and, therefore, reduced the consistency of responses between conditions, especially in regards to aggressive behavior. This finding highlights the importance of the particular stimulus dog chosen for the assessment and raises questions about the influence of the live stimulus dog on the results of the assessment. In our study, we used only a single real dog and a single fake dog. It is possible that test dogs may react differently to fake and real dogs of differing appearances and behaviors. Future research should consider using a single test dog with different real and fake stimulus dogs in order to investigate the effect of the stimulus dog (real and fake) further. This type of research would help to identify a stimulus dog dependent effect which raises the question of the validity of these types of assessments.

Substantial agreement between the two dog stimulus conditions was reached on the friendly trait. This may result due to several reasons. One could be the fact that we included 'tail wag' in our friendly trait. Tail wagging was a very frequently displayed behavior that could imply many things. It could be interpreted as friendly or as an arousal behavior. Moreover, the fact that in some cases tail wagging preceded aggressive behavior may indicate that tail wagging may not always signal a friendly intent. Another possible explanation could be that a dog may be more likely to react to a fake dog like it was a real dog at a distance compared to an up close interaction. For example, in the Barnard study (2012), the plastic fake dog elicited high and stiff postures in the dog aggressive group when it was placed 3 m apart from the test dogs but only a few subjects from the dog aggressive group showed an aggressive reaction upon contact (presumably because at a close distance, the test dog knows the plastic dog is fake). Moreover, the risk of displaying friendly behavior at a distance is minimal; there is time for the test dog to assess a conspecific's behavior and intention as the dogs approach and change its own behavior in reaction. In our study, although some dogs failed to react to the fake dog even in close proximity, 22% of dogs jumped on the fake dog and 9% displayed a play bow to the fake dog, suggesting that these dogs might have been misled by the fake dog appearance even upon close contact. The texture of the plush dog might mimic a real dog's fur, making the fake dog look more realistic than the plastic dog that was used in the Barnard study.

Only moderate agreement was reached on the fearful trait with more dogs being fearful of the fake dog. We believe that the fake dog's frozen posture and direct look might have elicited fear in some test dogs. The fake dog could also have been perceived as a frightening novel object by some test dogs. In several studies, mechanical toys were used to assess dogs' fear (King et al., 2003; Ley et al., 2007). In King et al. (2003) and Ley et al. (2007), mechanical cars were used since they have been found to reliably provoke avoidance behavior in other species. Ley et al. (2007) found that fearful dogs who did not go through the treatment program for fear spent less time near novel objects and approached the toy less frequently. Although the plush dog is different from mechanical toys, it could still have been perceived as a novel artificial object by some test dogs since it does not smell or move like a real dog and consequently it may have induced more fearful reactions in some dogs than the live stimulus dog.

In conclusion, the fake dog elicited a variety of different reactions in test dogs and seems to be a useful device to evaluate friendly behavior toward other canines. However, the aggressive trait showed only a slight degree of agreement between reactions to the real and the fake stimulus dogs. Shelters should be aware that the results of a behavior evaluation's dog-to-dog assessment can be different depending not only on whether a fake or real dog is used, but *which* fake or real dog is used. Although this study did not investigate the predictive validity of either the fake or real dog, it does raise the question of how much

influence the stimulus dog, live or fake, has on the test dog's reactions and highlights the need for more research.

Acknowledgments

This study was supported by a grant from The Stanton Foundation. The funder had no involvement with the conduct of this study, interpretation of the results, or writing and submission of this manuscript.

The authors thank Kim Melanson, Carol Ahearn, Laney Nee, and Christopher D'Arpino and his dog Ferris for helping with technical preparation and implementation of the study, Dr. Amy Marder for her professional advice, and Liz Fay for her great onsite support during the study implementation as well as all the staff at the Animal Rescue League of Boston and the Worcester Animal Rescue League for their support of the study.

References

Barnard, S., Siracusa, C., Reisner, I., Valsecchi, P., Serpell, J.A., 2012. Validity of model devices used to assess canine temperament in behavioral tests. Appl. Anim. Behav. Sci. 138 (1), 79–87.

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.

Capra, A., Marazzini, L., Albertini, M., 2009. Are pit bulls different? Behavioral evaluation within a rehabilitation program of ex-fighting dogs. J. Vet. Behav.: Clin. Appl. Res. 4 (2), 76.

Center for Shelter Dogs, 2013. Match-Up II Shelter Dog Rehoming Program, http://centerforshelterdogs.org/Home/DogBehavior/MatchUpII.aspx (10/16/2013).

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Collins, K., Reid, P., Martinez, L., 2012. Bred to fight: evaluating play in pit bull puppies from fighting lines. In: Proceedings of the AVSAB/ACVB Meeting, San Diego, CA, pp. 31–36.

D'Arpino, S., Dowling-Guyer, S., Shabelansky, S., Marder, A.R., Patronek, G.J., 2012. The use and perception of canine behavioral assessments in sheltering organizations. In: Proceedings of the 2012 American College of Veterinary Behaviorists/American Veterinary Society of Animal Behavior Veterinary Behavior Symposium, pp. 27–30.

De Palma, C., Viggiano, E., Barillari, E., Palme, R., Dufour, A.B., Fantini, C., Natoli, E., 2005. Evaluating the temperament in shelter dogs. Behaviour 143, 1313–1334.

Dowling-Guyer, S., Marder, A., D'Arpino, S., 2011. Behavioral traits detected in shelter dogs by a behavior evaluation. Appl. Anim. Behav. Sci. 130 (3), 107–114.

Gazzano, A., Mariti, C., Papi, F., Falaschi, C., Foti, S., Ducci, M., 2010. Are domestic dogs able to calm conspecifics by using visual communication? J. Vet. Behav.: Clin. Appl. Res. 5 (1), 28–29.

King, T., Hemsworth, P., Coleman, G., 2003. Fear of novel and startling stimuli in domestic dogs. Appl. Anim. Behav. Sci. 82, 45–64.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics, 159–174.

Leaver, S.D.A., Reimchen, T.E., 2008. Behavioural responses of *Canis familiaris* to different tail lengths of a remotely-controlled life-size dog replica. Behaviour 145 (3), 377–390.

Ledger, R.A., Baxter, M.R., 1997. The development of a validated test to assess the temperament of dogs in a rescue shelter. In: Proceedings of the 1st International Conference on Veterinary Behavior Medicine, Birmingham, UK, pp. 87–92.

Ley, J., Coleman, G.J., Holmes, R., Hemsworth, P.H., 2007. Assessing fear of novel and startling stimuli in domestic dogs. Appl. Anim. Behav. Sci. 104 (1–2), 71–84.

Mackinnon, A., 2000. A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. Comput. Biol. Med. 30 (3), 127–134.

Marder, A., Ahearn, C., D'Arpino, S., Dowling-Guyer, S., Fantuzzi, J., Johnston, N., MacDougall, L., Melanson, K., Patronek, G., Shabelansky, A., Woo, C., 2013. Development and implementation of a unique online portal for collecting data from a standardized behavior evaluation with shelter dogs. J. Vet. Behav.: Clin. Appl. Res. 8 (4), e40–e41.

Mariti, C., Papi, F., Ducci, M., Sighieri, C., Martelli, F., Gazzano, A., 2010. Domestic dogs display calming signals more frequently towards unfamiliar rather than familiar dogs. J. Vet. Behav.: Clin. Appl. Res. 5 (1), 62–63.

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

Reid, P., Collins, K., 2012. Assessing Conspecific Aggression in Fighting Dogs. In: Proceedings of the AVSAB/ACVB meeting, San Diego, CA, pp. 37–39.

Siegel, S., 1956. Nonparametric Statistics: For the Behavioral Sciences. McGraw-Hill Book Company Inc., New York, pp. 116–127.

Svartberg, K., 2007. Individual differences in behavior–dog personality. In: Jensen, P. (Ed.), The Behavioral Biology of Dogs. CABI, Oxfordshire, UK, pp. 151–167.

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

Uebersax, J., 2013. The Myth of Chance-Corrected Agreement, http://www.john-uebersax.com/stat/kappa2.htm (9.11.2013).

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

Weiss, E., 2007. Meet Your Match SAFERTM Manual and Training Guide, http://www.aspcapro.org/safer