Do you see what I see? Can non-experts with minimal training reproduce expert ratings in behavioral assessments of working dogs?

ARTICLE *in* BEHAVIOURAL PROCESSES · SEPTEMBER 2014

Impact Factor: 1.57 · DOI: 10.1016/j.beproc.2014.09.028

6 AUTHORS, INCLUDING:

Jamie L. Fratkin
University of Texas at Austin
3 PUBLICATIONS **33** CITATIONS
SEE PROFILE

David L. Sinn
University of Tasmania
26 PUBLICATIONS **776** CITATIONS
SEE PROFILE

Samuel D Gosling
University of Texas at Austin
124 PUBLICATIONS **10,859** CITATIONS
SEE PROFILE

Behavioural Processes 110 (2015) 105–116

Contents lists available at ScienceDirect

Behavioural Processes

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

Do you see what I see? Can non-experts with minimal training reproduce expert ratings in behavioral assessments of working dogs?

Jamie L. Fratkin[a,*], David L. Sinn[a,b], Scott Thomas[c], Stewart Hilliard[d], Zezelia Olson[a], Samuel D. Gosling[a]

[a] The University of Texas at Austin, Department of Psychology, 108 E. Dean Keeton Stop A8000, Austin, TX 78712, USA
[b] University of Tasmania, School of Zoology, Private Bag 5, Hobart, Tasmania 7001, Australia
[c] Transportation Security Administration Canine Breeding and Development Center, 341 TRS/TSA Puppy Program, 1320 Truemper Street, Bldg 9122 Suite 2 Room 2701, Lackland AFB, TX 78236, USA
[d] Department of Defense Military Working Dog Breeding Program, 341 TRS, 9225 Truemper Street, Bldg 9225, Lackland AFB, TX 78236, USA

ARTICLE INFO

Article history:
Available online 28 September 2014

Keywords:
Dog
Behavior assessment
Reliability
Validity
Dog experience

ABSTRACT

Working-dog organizations often use behavioral ratings by experts to evaluate a dog's likelihood of success. However, these experts are frequently under severe time constraints. One way to alleviate the pressure on limited organizational resources would be to use non-experts to assess dog behavior. Here, in populations of military working dogs (Study 1) and explosive-detection dogs (Study 2), we evaluated the reliability and validity of behavioral ratings assessed by minimally trained non-experts from videotapes. Analyses yielded evidence for generally good levels of inter-observer reliability and criterion validity (indexed by convergence between the non-expert ratings and ratings made previously by experts). We found some variation across items in Study 2 such that reliability and validity was significantly lower for three out of the 18 items, and one item had reliability and validity estimates that were impacted heavily by the behavioral test environment. There were no differences in reliability and validity based on the age of the dog. Overall the results suggest that ratings made by minimally trained non-experts for most items can serve as a viable alternative to expert ratings freeing limited resources of highly trained staff.

This article is part of a Special Issue entitled: Canine Behavior.

1. Introduction

Working dogs are engaged in a wide variety of tasks that improve the lives of humans, ranging from guiding the visually impaired to detecting roadside bombs in combat zones. Programs devoted to breeding and training working dogs often evaluate the dogs' behaviors at several ages to determine their likelihood of success and to evaluate their performance. Experts, usually professionals with years of experience working with and training dogs, are typically called upon to make such behavioral ratings. These same experts are also often required to perform many other program-related tasks, which may leave them with little time to make ratings. If experts' ratings could be reproduced by non-experts (e.g., members of the public or selected groups of volunteers), then the experts would be free to engage in other tasks and ratings could be undertaken on a much larger scale, potentially improving the efficiency of working dog programs. Here we evaluate whether behavioral ratings of working-dog behaviors in behavioral assessments made by minimally trained non-experts can be reliable, and if so, whether non-expert ratings faithfully reproduce the ratings made by experts rating the same dogs. Specifically, in two independent working-dog programs, we examined the reliability of non-expert ratings in terms of inter-observer reliability, and we examined criterion validity in terms of the convergence between the non-experts' ratings and the expert ratings they are designed to reproduce.

In general, literature examining similarities and differences between expert and non-expert behavioral ratings in animals has yielded somewhat contradictory findings. For example, two studies, one with dogs (Tami and Gallagher, 2009) and one with pigs (Wemelsfelder et al., 2012) found that behavioral ratings made by people with different types of animal expertise did not vary significantly across the different groups of people, suggesting that experts and non-experts interpreted the behaviors similarly. Other research found that for some items, dog owners rated their dogs differently than people with different types and levels of dog experience (trainers, other dog owners, and non dog owners; Mirkó et al., 2013). However, in another study of pigs, inter- and

* Corresponding author. Tel.: +1 3019083006.
E-mail addresses: fratkijl@utexas.edu, fratkijl@gmail.com (J.L. Fratkin).

http://dx.doi.org/10.1016/j.beproc.2014.09.028
0376-6357/

intra-observer reliability among non-expert raters was higher than inter and intra-observer reliability among expert raters (Cloudard et al., 2011). In one study of zoo chimpanzee behavior, researchers found that ratings from non-experts did not accurately reproduce expert ratings (Duncan and Pillay, 2012) and in another study, people with more dog experience were better at rating fear in dogs than people with less dog experience (Wan et al., 2012). Finally, another study of non-human primates found that inter-observer reliability was lower for raters less familiar with individual primates than it was for observers more familiar with them (Martau et al., 1985). In short, the literature fails to provide a clear answer to the key practical question of whether non-experts can be safely used in place of experts for behavioral ratings of working dogs.

Fortunately, despite the lack of clarity emerging from past studies, the broader research literature does offer some clues regarding the factors that might affect the reliability and validity of behavioral ratings given during behavioral assessments of working dogs. For example, some items and behaviors are more reliably judged than are others in humans and other non-human animals (Gosling, 2001; Gosling et al., 1998; John and Robins, 1993). In one literature review, Gosling (2001) examined inter-observer reliability across many non-human animal species and found items related to Extraversion to be high in inter-observer reliability, whereas items related to Agreeableness were low in inter-observer reliability. Factors such as how visible the items are, how evaluative (or socially desirable) they are, and their frequency can all affect the reliability and validity of behavioral measurement (Gosling et al., 1998; John and Robins, 1993). In particular, visible behaviors like talkativeness tend to be judged more reliably than less visible behaviors like daydreaming and non-evaluative behaviors such as talkativeness tend to be judged more reliably than more evaluative behaviors like ignorance (Funder and Dobroth, 1987; John and Robins, 1993). In the working-dog domain, some stimuli might elicit behaviors that are easy to observe (e.g., fear) whereas other stimuli might be associated with much more subtle behavioral indicators (e.g., anxiety); as a result inter-observer reliability could differ across items depending on how easily the items are observed (Bahlig-Pieren and Turner, 1999; Tami and Gallagher, 2009). In addition, there might be some behaviors that can only be properly interpreted after extensive experience with the species. One widely known example from the primate literature is the "fear grin" in chimpanzees which signals fear but on the basis of its superficial similarity to the human smile is often interpreted by non-experts as signaling happiness (Waller and Dunbar, 2005).

Another factor potentially affecting reliability and validity of behavioral measures is the age of the target subject. Research in humans and non-human animals (including dogs) has shown that behavioral consistency in many instances tends to increase with age (e.g., Fratkin et al., 2013; Sinn et al., 2008). In humans, research suggests that individuals who are more consistent in their behavior are more judgeable (Funder, 1995). Following this logic, inter-observer reliability and validity may be higher in older dogs than in younger dogs. Determining whether reliability and validity is impacted by the age of the dogs could have significant practical applications for working-dog programs because it could throw light on the optimal age at which the dogs should be rated.

1.1. Behavioral codings and behavioral ratings

In behavioral assessments for working-dog programs, dogs are typically exposed to standardized stimuli or situations and the dogs' observed behavioral responses are recorded using either behavioral codings or behavioral ratings (Fratkin et al., 2013; Jones and Gosling, 2005). Behavioral codings are designed to capture observed, discrete behaviors, and typically use frequency counts and/or durations (e.g., the number of times a dog crosses a grid-line marked on the floor, or the total time spent moving during an assessment). Behavioral ratings consist of broader subjective judgments regarding observed dog behavior during an assessment (e.g., the dog's level of confidence; Gosling, 2001). Both methods have been shown, under some circumstances, to attain acceptable levels of inter-observer and test-retest reliability (Fratkin et al., 2013; Jones and Gosling, 2005). Here we focus on behavioral ratings because they are widely used in working-dog programs (e.g., Maejima et al., 2007; Paroz et al., 2008; Sinn et al., 2010) and were the method already in use in the two working-dog programs to which we had access.

1.2. The present research

In two studies, we examined reliability and validity of behavioral ratings made by minimally trained non-expert raters in two working-dog populations located at Lackland Air Force Base in San Antonio, Texas, USA. In Study 1, we examined the inter-observer reliability and criterion validity of non-expert raters (using the expert ratings as the validity criterion) in a population of Military Working Dogs (MWDs). In Study 2, we examined the inter-observer reliability and criterion validity of non-expert raters in a population of working dogs bred and trained by the USA's Transportation Security Administration Canine Breeding and Development Center (TSA-CBDC) as explosive detection dogs. We were particularly interested in whether some kinds of behaviors were associated with higher reliability and validity than were others (Studies 1 and 2) and whether the age of the dog influenced reliability and validity (Study 2).

2. Methods

2.1. Study 1: canine subjects

A total of 25 2-month old Belgian Malinois (13 males, 12 females) bred and reared as part of the USA's 341st Training Squadron (TRS) at Lackland Air Force Base, San Antonio, Texas participated in Study 1. Prior to behavioral ratings, dogs from the same litter had lived together in a kennel with a grassy enclosure and with objects with which they could play. Dogs interacted with caretakers on a daily basis and were encouraged to play with balls and rubber toys and were regularly exposed to novel noises and objects in their kennel environment. Puppies were weaned between 5 and 6 weeks of age.

2.2. Study 1: behavioral assessments

Dogs' behaviors were rated in three different domains, called 'toy interest', 'environment', and 'bite work'. Each of the three domain assessments lasted for approximately 5 min, and behavior in each of the domains was measured on the same day for a given dog. All dogs were given the three assessments in the same order (as described below) in a room indoors. Dogs were assessed off-leash for the entire assessment, and two to three people (a trained handler, a camera operator, and often an assistant) were present during the assessment. Assessments were videotaped and an expert (a different person from the handler) rated the dog on a separate occasion on 10 items based on observed behaviors from videotape recordings (see Fig. 1). Each rating was recorded on a continuous scale and dogs could receive any score on the scale's continuum, including fractional points. Ratings were made on a 5-point Likert scale with a '5' representing a high score and a '1' representing a low score (see Online Supplementary Materials Table 1 for a thorough description of each item). The behavioral assessments at 2-months of age were not used in this program as the basis for accepting or rejecting dogs from training.

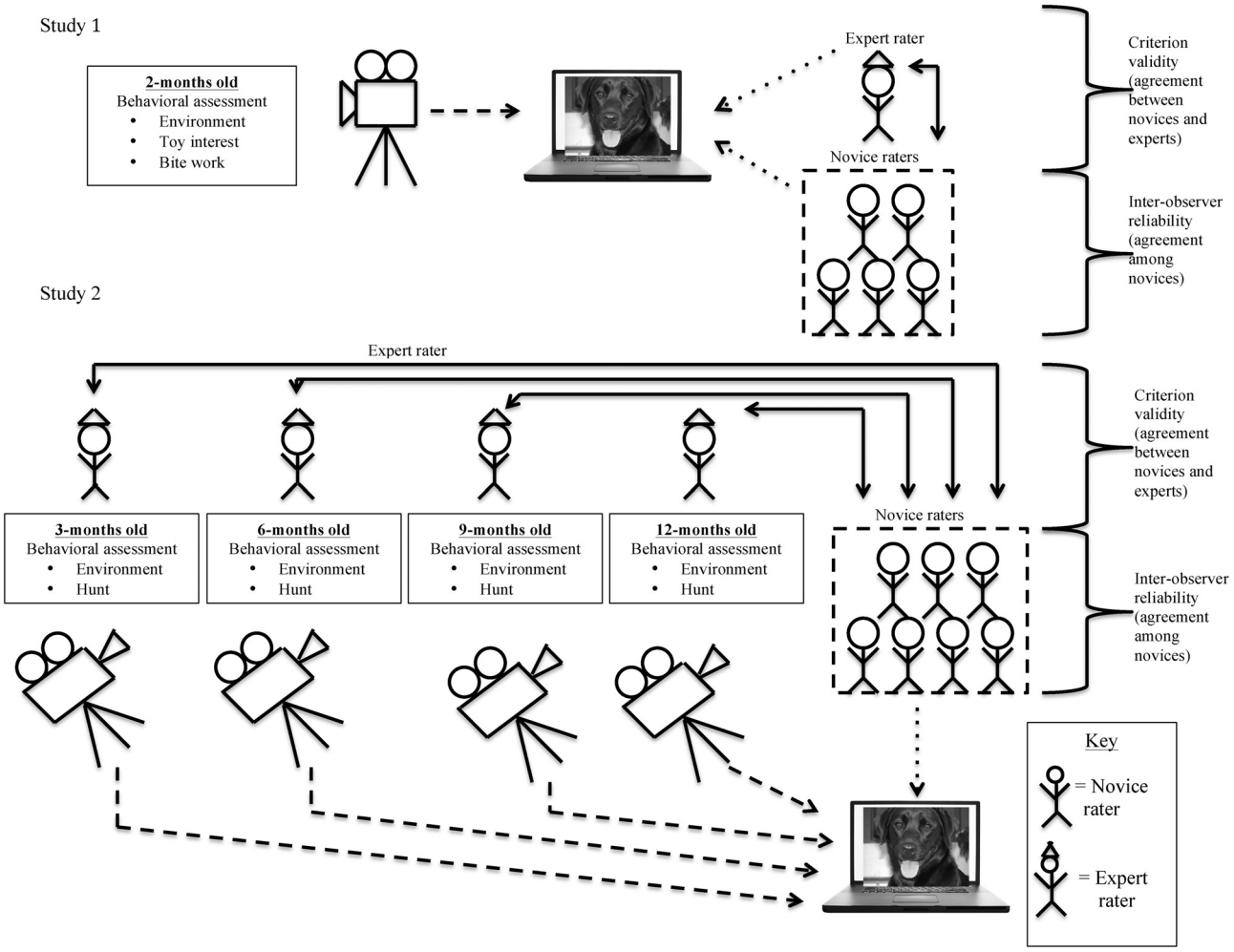

Fig. 1. Overview of study design for Studies 1 and 2.

2.2.1. Toy interest

The toy interest domain assessed a dog's reaction to two objects (a sock and a ball). First, the handler threw a ball several times and rated the dog's object pursuit (the speed and intensity by which the dog chases after the object), object possession (the physical possession of the object), and object interest (the dog's continued visual fixation and physical orientation toward the object). The handler repeated the same assessment using a sock to assess the dog's overall pursuit, possession, and interest. The final ratings for object possession, object pursuit, and object interest were an average of the ratings taken from the two object assessments (one with the ball and one with the sock).

2.2.2. Environment

In the environment assessment domain, the dog was exposed to environmental stimuli that had the potential to elicit a fear response. First, dogs were presented with a vacuum cleaner, an object that was noisy and novel to the dogs. Dogs were placed in the center of the room with some food. The handler initially turned on a vacuum cleaner that was near a wall on the side of the room and waited to see if the dog approached the vacuum cleaner. If the dog did not approach the vacuum cleaner, the handler turned the vacuum cleaner off and the dog was given the opportunity to approach the vacuum cleaner. If the dog still did not approach the vacuum cleaner, the handler encouraged the dog to approach the vacuum cleaner by enthusiastically pointing to the object and calling to the dog. After about a minute, this part of the assessment was concluded. Next, dogs were presented with a metal can filled with coins, the "noisy can", which dogs had not previously encountered. Dogs were again placed in the center of the room with food, and the noisy can was dropped on the floor on the side of the room. The handler waited to see if the dog approached the noisy can, and if the dog did not approach the can, the handler encouraged the dog to approach the can by enthusiastically pointing to the object and calling to the dog. Again, after about a minute, this part of the assessment was concluded. The final part of the environment assessment examined the dog's avoidance of a chair (on wheels) and a vacuum cleaner. The dog was placed in the center of the room with a bowl filled with food. While the dog was eating, the handler first rolled a chair toward the dog. After the chair part of the assessment, the chair was removed and the dog was again placed in the center of the room with a bowl filled with food. Then, the handler rolled a vacuum cleaner that was turned off toward the dog. The dog's reaction to the chair and the vacuum cleaner were assessed separately and then averaged to create an 'approach avoidance' rating. Dogs were rated on three items during the environment assessment (vacuum approach, noisy can approach, and approach avoidance).

2.2.3. Bite work

For the bite work assessment domain, the dog was shown a rag and was encouraged to bite and hold onto it. The dog's bite

interest (the dog's physical and visual orientation to the rag), bite commitment (the dog's movement toward the rag), bite quality (the strength of the dog's biting of the rag), and bite steadiness (the constancy of the dog's bite to the rag) were all rated during this assessment. A dog scoring a '5' on bite interest had very high energy in attempting to bite and touch the rag and attended to every movement of the rag, while a dog scoring a '1' on bite interest was disinterested, inhibited, or preoccupied with soliciting attention from the handler. A dog scoring a '5' on bite commitment vigorously attempted to seize the rag, while a dog scoring a '1' on bite commitment may have followed the rag for a moment, but was mostly disinterested in the rag. A dog scoring a '5' on bite quality bit the rag forcefully with most or all of his/her mouth during the entire test, while a dog scoring a '1' on bite quality either did not bite the rag at all or bit the rag briefly. A dog scoring a '5' on bite steadiness did not release the rag under any circumstance and showed little or no change in biting the rag through the test, while a dog scoring a '1' on bite steadiness never bit the rag.

2.3. Study 1: expert and non-expert behavioral ratings

Each dog was rated after the assessment by a single observer; this rater was labeled as an "expert" because he had 22 years of experience rating similar dog assessments and had rated 39 dogs using this exact dog assessment. "Non-expert raters" were psychology undergraduate students at the University of Texas at Austin who volunteered or worked as research assistants. Non-expert raters consisted of four females and one male, aged 19–23, and were of varied ethnic origin (Caucasian, African American, mixed race, and Hispanic). None of the non-expert raters had previous experience in rating dog behavior, but all of them had lived with or were currently living with dogs for a period of time ($M = 12.10$ years, SD = 6.48 years).

2.4. Study 1: non-expert rater training

Non-experts were trained individually. Non-experts were first given a standardized form that defined the items and gave examples of observed behaviors that corresponded to different ratings for each item; the first author also gave a verbal explanation of what each item meant. Then the non-experts were given a training video, which consisted of six video segments, with each segment depicting a dog going through the tasks described above. Non-experts were shown the first video segment and told what ratings the expert had assigned to the dog on each of the 10 items. The non-experts were allowed to ask questions if they were unsure about what a particular item meant and were allowed to rewind, pause, and re-watch the training video as much as they wanted. The non-experts continued to watch the remaining five training video segments. After watching each segment, they gave ratings for each item on the dog in the video, and after watching all five video segments they were told what ratings the expert had assigned to each dog on each of the 10 items. If non-experts had a majority of their responses (51%) within one number of the expert scores for all 10 items they were allowed to move on to the videos from the study. All non-experts reached this criterion after watching the five videos. The training took approximately 1 h.

For data collection, the non-expert raters watched all the dog videos, and rated each dog using the standardized form. Non-expert raters viewed videos at their own pace and completed rating them in approximately 2–5 months. The non-experts viewed the videos in assigned random orders. The non-experts were instructed to make their scores independently and not to talk to any other non-expert about their ratings during the study period. Non-experts were allowed to pause and rewind and re-watch the videos as much as they desired.

2.5. Study 1: data analysis

Statistical analyses were performed using SPSS version 20 (SPSS Inc., Chicago, IL, USA). To evaluate inter-observer reliability of the non-expert ratings, we computed the intra-class correlation coefficient (ICC [2,1]; Shrout and Fleiss, 1979) for each rating given by non-expert raters and compared it to the 0.70 threshold sometimes proposed as a threshold for acceptable reliability (Cicchetti, 1994; but see John and Soto, 2007). To evaluate the criterion validity of the non-expert ratings, the non-experts' single and aggregated scores were correlated with the expert scores. Single scores were calculated by estimating the correlation between the expert's score and each individual non-expert's score; the average of each ratings' correlation coefficient for each non-expert was then computed using Fischer's r to z transformation. Criterion validity was also examined using aggregate measures, which were calculated by first taking an average of the five non-expert score and correlating this aggregated non-expert score with the expert's score. Pearson correlations were used for all validity estimates. We present both single and aggregate estimates in tables, but report only single measures in the text; single measures agreement coefficients are not inflated by the effects of aggregation (Epstein, 1983). Significant differences in inter-observer reliability and validity estimates between items were assessed graphically by comparing estimates and (non) overlapping 95% confidence intervals (Cumming et al., 2007).

2.6. Study 2: canine subjects

Subject dogs were a total of 51 unique purpose-bred detection dogs (27 males, 24 females) located at the Lackland Air Force Base in San Antonio, Texas as part of the TSA-CBDC program. Most dogs ($N = 41$) were fostered and raised outside of the kennels for some time of their development, but some dogs were raised solely in a kennel environment ($N = 10$). Dogs were a mix of Labrador Retrievers ($N = 40$), Vizslas ($N = 6$), and Labrador/German Shorthaired Pointer crosses ($N = 5$).

2.7. Study 2: behavioral assessments

Dogs were rated in two different behavioral assessment domains (an 'environment' assessment domain and a 'hunt' assessment domain) at 3-months, 6-months, 9-months, and 12-months of age. However, due to logistical constraints, some dogs were rated more than once (i.e., they contributed data at two different ages), so sample sizes varied. Thus, from the total pool of 51 dogs, for the environment test, 17 were assessed at 3-months (14 Labrador Retrievers, 3 Labrador/German Shorthaired Pointer crosses), 22 at 6-months (13 Labrador Retrievers, 4 Vizslas, 5 Labrador/German Shorthaired Pointer crosses), 23 at 9-months (13 Labrador Retrievers, 5 Vizslas, 5 Labrador/German Shorthaired Pointer crosses), and 17 at 12-months (9 Labrador Retrievers, 3 Vizslas, 5 Labrador/German Shorthaired Pointer crosses). For the hunt test, 26 were assessed at 3-months (16 Labrador Retrievers, 5 Vizslas, 5 Labrador/German Shorthaired Pointer crosses), 25 at 6-months (15 Labrador Retrievers, 5 Vizslas, 5 Labrador/German Shorthaired Pointer crosses), 24 at 9-months (14 Labrador Retrievers, 5 Vizslas, 5 Labrador/German Shorthaired Pointers crosses), and 18 at 12-months (15 Labrador Retievers, 3 Vizslas). The same two people handled all dogs during all behavior assessments. The environment assessment lasted for approximately 10–20 min. The hunt assessment lasted for approximately 5–10 min. Both assessments at each age were performed on the same day and were videotaped by either a head camera (a camera attached to the handler's head)

Table 1
Operational definitions of itemsassessed during the Transportation Security Administration Canine Breeding and Development Centerbehavioral assessments.

Domain	Item	Description
Environment	Confidence	An environmentally conditioned acceptance of safety, in other words, a measure of the degree of fear or lack thereof (high fear = low confidence). Typical signs of fear include extended attention, freezing, and avoidance. A '1' is given if subject dog froze in response to a stimulus, a '5' is given if the dog displayed no fear.
Environment	Concentration	The dog's focus on the area of search in general, and is a measure of the degree of distraction during the assessment. Typical signs of distraction are objects (sticks, leaves, etc.) on the ground, people, self, and leash biting. A '1' is given if the subject dog is continually distracted, a '5' is given if the dog displayed complete focus.
Environment	Responsiveness	The dog's ability to react to corrections or encouragement from the handler. Typical signs include a lack of response to correction or to encouragement (verbal and physical praise). A '1' is given if subject dog shows no reaction to handler's corrections or encouragements, a '5' is given if the dog responded to all handler corrections and encouragement.
Environment	Initiative	A dog's willingness to walk at the end of the leash and investigate the environment on his or her own without being asked by the handler. A '1' is given if the subject dog lags behind on the leash and refuses to initiate investigation, a '5' is given if the dog investigates the environment on his or her own in multiple instances.
Environment	Excitability	The dog's enthusiasm during the walk, which is considered to be either more or less than appropriate. Over excitability typically displayed by a dog jumping on everything, even when not asked to. Under excitability is typically displayed by a lackadaisical approach during the assessment. A '1' is given if the subject dog exhibits too little or too much enthusiasm, a '5' is given if the dog exhibits an appropriate level of enthusiasm.
Environment	Hearing Sensitivity	The dog's sensitivity to noises. Hearing sensitivity is a measure of if a dog reacts to any sound stimulus. A '1' is given if the subject dog exhibits either too little or too much of a reaction to noise, a '5' is given if the dog exhibits an appropriate reaction to noise.
Environment	Body Sensitivity	The dog's physical sensitivity to touch by the handler. Body sensitivity is a measure of if a dog reacts to touch, praise, or correction. This may be displayed by fearfulness after praise or corrections. A '1' is given if the subject dog exhibits either too little or too much of a reaction to touch, a '5' is given if the dog exhibits an appropriate reaction to touch.
Environment	Chase Retrieve	The speed at which the dog runs for a toy and the dog's desire to pick up the toy. A '1' is given if the subject dog shows no desire to pick up or chase after the toy, a '3' is given if the dog chases after and picks up the toy quickly.
Environment	Independent Possession	The desire to possess the toy independently of the handler. A '1' is given if the subject dog shows no desire to continually possess the toy, a '3' is given if the dog continuously keeps the toy in possession.
Environment	Physical Possession	The dog's desire to hold on to a toy playing tug-of-war with the handler. This is measured by the dog's grip on the towel. A '1' is given if the subject dog either has no grip on the towel or allows the handler to take the towel without too much effort, a '3' is given if the dog holds a strong grip on the towel, which makes it difficult for the handler to take away the towel.
Hunt	Chase Retrieve	The speed at which the dog runs for a toy and the dog's desire to pick up the toy. A '1' is given if the subject dog shows no desire to pick up or chase after the toy, a '5' is given if the dog chases after and picks up the toy quickly.
Hunt	Physical Possession	Measured based on the dog's grip on the towel. A '1' is given if the subject dog either has no grip on the towel or allows the handler to take the towel without too much effort, a '5' is given if the dog holds an extremely strong grip on the towel in which the handler has difficulty in taking the towel away.
Hunt	Hidden Grass	A '1' is given if the subject dog has extreme trouble finding a towel hidden in grass and the handler has to point out the location to the dog or the dog never finds the towel, a '5' is given if the dog finds the towel without help, using olfaction.
Hunt	Independent Possession	The desire to interact and possess the toy independently of the handler. A '1' is given if the subject dog shows no desire to continually possess the towel, a '5' is given if the dog continuously keeps the towel in possession.
Hunt	Hidden 1	A '1' is given if the subject dog cannot find the towel without the handler pointing it out or if the dog never finds the towel, a '5' is given if the dog finds the towel without help, using olfaction.
Hunt	Hidden 2	A '1' is given if the subject dog cannot find the towel without the handler pointing it out or if the dog never finds the towel, a '5' is given if the dog finds the towel without help, using olfaction.
Hunt	Mental Possession	This is evaluated by the dog's ability to focus on the hiding of the towel. A '1' is given if the subject dog does not focus on the towel at all, a '5' is given if the dog shows complete focus on the towel.
Hunt	Activity	The dog's ability to use his/her energy appropriately. A '1' is given if the subject dog does not use his/her energy appropriately at all, a '5' is given if the dog uses his/her energy appropriately.

or a handheld camera held by an assistant. The environment assessment was given first to 9-month and 12-month old dogs; the hunt assessment was given first to 3-month and 6-month old dogs. A total of 10 items were rated by the two expert handlers in the environment assessment, and a total of eight items were rated in the hunt assessment (Table 1). Each rating was given on a Likert scale in which only full points could be given. Experts rated dogs immediately after the assessments were given, whereas non-experts rated dogs at a later date after watching the videos of the assessments. Videos were labeled by a video number, rather than the dog's name, so raters were blind to the dog's identity on the video, and did not know when the same dog was assessed at another age. The behavioral assessments in this program were not used as the basis for accepting or rejecting dogs from training.

2.7.1. Environment

In the environment assessment domain, each dog was rated by one of the handlers. The first part of the environment assessment consisted of the handler walking the dog through a novel environment, which was different depending on the age of the dog. The walk was standardized and designed to allow the dog to encounter a range of stimuli and situations. The handler rated the dog on seven items immediately after the environment assessment: confidence, concentration, responsiveness, initiative, excitability, hearing sensitivity, and body sensitivity (see Table 1 for detailed descriptions of these items). For each of the items, the handler rated dogs on a 5-point Likert scale. A '1' indicated a lower expression of an item and a '5' indicated a higher expression of an item.

The second part of the environment assessment involved rating the dog's reaction to a toy. After walking and observing dogs for approximately 7–15 min, the handlers stopped walking, removed a hidden toy, and threw the toy a few meters in front of the dog. Handlers rated the dogs on how intensely the dog ran after the toy (chase retrieve). Next, the handler initiated a tug-of-war game with the dog and rated how hard the dog gripped the toy (physical possession). Finally, the handler gave the toy back to the dog after the tug-of-war game and walked with the dog back to the starting location. Handlers then rated dogs on whether the dog kept the toy in his/her mouth on the walk back to the truck (independent possession). Chase retrieve, independent possession, and

physical possession were all rated on a 3-point Likert scale, with a '1' indicative of dogs that generally lack expression of these items, and a '3' indicative of dogs that expressed high levels of these items.

The environment assessment location depended on the dog's age. Dogs that were 3-months of age were assessed in a general store that included aisles, shelves, boxes, grates, and treadmills for dogs to explore. Dogs that were 6-months of age were assessed at a woodshop, which had unexpected noises, machinery, various rooms and floors with differences in lighting, as well as small spaces and stairs. Dogs that were 9-months of age were assessed in an airport cargo area, which had cargo trucks and trailers that dogs could jump on and into, as well as unusual stairs and loud noises. Finally, dogs that were 12-months of age were assessed at an airport terminal, in which they were exposed to the baggage claim area, the airport parking lot, and the entrance to the airport.

2.7.2. Hunt

In the hunt assessment domain, both handlers worked together to rate one dog at a time. Dogs were always rated in the same location in an open field regardless of age. In this assessment, one of the handlers initiated play with the dog using an odor-saturated rolled cotton towel, and then threw the towel into an open area 6–9 meters away from the dog (or about 3 meters for 3-month old dogs). The handler waited for the dog to retrieve the towel and then initiated another game of tug-of-war. Next, the handler took the towel from the dog and threw the towel into tall grass so that the towel was visually hidden. If the dog had difficulty finding the towel, the second handler stepped in to the grass to help assist the dog. During the 'hidden grass' portion of the hunt assessment domain, the handlers specifically noted whether the dog used olfaction, and not vision, to retrieve the towel.

After the dog retrieved the towel from the tall grass, the dog was allowed to retain the scented towel while the handlers walked to a part of the open field where 10 overturned plastic flowerpots had been placed in a straight line. As the dog moved to the line of flowerpots, the handlers rated the dog's continued possession of the towel. Once the dog and handlers reached the flowerpots, one handler took the towel away from the dog, gave it to the second handler, and held the dog at one end of the flowerpots. The second handler walked the entire line of flowerpots, and made physical contact with each pot, but placed the towel under only one pot so that it was visually hidden from the dog. In the first flowerpot trial, the second handler stopped at the opposite end of the flowerpot line, turned with outstretched arms and open hands, and faced the first handler and dog. The dog was then released and allowed to search for the towel without any further cues. Once the dog retrieved the towel a handler played tug-of-war with the dog as a reward. Next, the flowerpots were replaced, and a second trial was conducted. The second trial was identical to the first with the sole exception that after walking the entire line of flowerpots and hiding the towel, the second handler returned to the place where the first handler and dog were located prior to the dog being released to search.

Ratings in the hunt assessment domain were given according to the dog's chase retrieve, physical, mental, and independent possession of the towel, the dog's ability to find the towel underneath the empty flowerpot containers (hidden 1, hidden 2), and the dog's overall activity level (Table 1). For each of the items, dogs were rated on a 5-point Likert scale, with lower scores indicating a lack of expression of a particular item, and higher scores indicated a high degree of expression of a particular item. All behavioral ratings were given at the end of the hunt assessment, and were thus based on global responses observed during the whole assessment. In contrast to the environment assessment, both expert handlers discussed and agreed to ratings given to dogs in hunt assessments.

2.8. Study 2: expert and non-expert behavioral ratings

The two TSA-CBDC handlers (with 11 and 9 years of experience rating these assessments) were considered our "experts" in Study 2 and provided ratings for dogs immediately after the assessments. In addition, 21 human participants serving as our "non-experts" rated dogs from videotapes. Of the non-experts, 17 were complete novices in that they had no experience other than living with a dog; four participants had at least three years of experience beyond living with a dog (e.g., dog-training experience or dog-research experience; see Supplementary material for details of the analysis between complete novices and novices with some experience). Each dog was rated by a total of seven minimally trained non-experts (five complete novices and two with some dog experience). Non-experts were classified as such because they had no formal training in rating explosive detector-dog behavioral assessments. Non-experts were 18 females and three males, aged 19–27, of different ethnic origins (Caucasian, Hispanic, African American, Asian American, and mixed race).

2.9. Study 2: non-expert rater training

Non-experts were trained individually. Non-experts were first given a standardized form that defined all of the items in the environment assessment domain; the first author also gave a verbal explanation of what each rating meant in the environment assessment domain. Then the non-experts were given a training video, which consisted of four video segments, with each segment depicting a dog going through the environment assessment. Non-experts were shown the first video segment and told what ratings the expert had assigned to the dog for each of the 10 items. The non-experts were allowed to ask questions if they were unsure about what a particular item meant and were allowed to rewind, pause, and re-watch the training videos as much as they wanted. The non-experts continued to watch the remaining three training video segments. After watching each segment, they gave ratings to the dog in the video, and after watching all three video segments they were told what ratings the expert had assigned to each dog on each of the 10 items. Non-experts were then given a standardized form that defined all of the items in the hunt assessment domain; the first author also gave a verbal explanation of each rating in the hunt assessment domain. Non-expert raters then were given a training video, which consisted of six hunt training segments and went through the same process as they did for the environment assessment. Separately, for each assessment domain (environment and hunt) if a majority of their responses (51%) were within one number of the expert scores in both assessment domains they were allowed to move on to the videos from the study. All non-expert raters met this criterion after watching the 10 training videos. The training session took approximately 1 h.

For data collection, the non-expert raters watched all the dog videos, and rated each dog using the standardized form. Non-expert raters viewed videos at their own pace and in assigned random orders. The non-experts were instructed to make their scores independently and not to talk to any other non-expert about their ratings during the study period. Non-experts were allowed to pause and rewind and re-watch the videos as much as they desired.

2.10. Study 2: data analysis

Statistical analyses were performed using SPSS version 20 (SPSS Inc., Chicago, IL, USA). The non-experts watched videos from the study for the dogs at 3-months, 6-months, 9-months, and 12-months of age in a randomized order that was different for each rater. In addition, for each individual video, five non-experts watched the videos a second time to assess intra-observer

Table 2
Intra-class correlation coefficients (ICCs) and their confidence intervals (CIs), and Pearson correlations (*r*) between non-experts means (single and aggregate) and expert ratings for each item given during the military working dog behavioral assessment (Study 1) using 5 non-expert raters. ICCs measure inter-observer reliability and r measures criterion validity. Aggregate estimates are listed in parentheses for Pearson correlations.

	ICCs (95% CIs)	*r*	95% CI lower	95% CI upper
Object possession	0.99 (0.98–1.00)	0.57 (0.59)	0.23 (0.25)	0.79 (0.80)
Object pursuit	0.98 (0.96–0.99)	0.90 (0.93)	0.78 (0.85)	0.96 (0.97)
Object interest	0.98 (0.96–0.99)	0.96 (0.98)	0.91 (0.95)	0.98 (0.99)
Vacuum approach	0.89 (0.80–0.95)	0.64 (0.76)	0.32 (0.51)	0.83 (0.89)
Noisy can approach	0.94 (0.89–0.97)	0.83 (0.92)	0.64 (0.82)	0.92 (0.97)
Approach avoidance	0.98 (0.96–0.99)	0.89 (0.93)	0.76 (0.84)	0.95 (0.97)
Bite interest	0.92 (0.85–0.96)	0.75 (0.86)	0.50 (0.70)	0.89 (0.94)
Bite commitment	0.94 (0.90–0.97)	0.70 (0.77)	0.42 (0.54)	0.86 (0.89)
Bite quality	0.94 (0.88–0.97)	0.78 (0.84)	0.56 (0.67)	0.90 (0.93)
Bite steadiness	0.93 (0.87–0.97)	0.83 (0.91)	0.65 (0.80)	0.92 (0.96)
Average	0.96 (0.91–0.98)	0.82 (0.89)	0.62 (0.76)	0.92 (0.95)

reliability. Non-experts watched videos a second time in varying intervals from the first video, anywhere from less than a day later to 117 days later. The results for intra-observer reliability are presented in the Supplementary materials.

To evaluate reliability for each behavioral rating, we used the intra-class correlation coefficient (ICC [1,k]; Shrout and Fleiss, 1979) and compared it to the 0.70 threshold sometimes proposed as a threshold for acceptable reliability (Cicchetti, 1994; but see John and Soto, 2007). For some ratings at some ages, there was no variability in ratings. In other words, all raters gave the same rating for the same item across all dogs at a particular age (e.g., only 3's were given by all non-experts raters for all 9-month old dogs for the item 'hearing sensitivity'). The result of this lack of variability is that ICCs cannot be computed for these item/age combinations. Yet the fact that all raters put down the same number indicates high agreement. Therefore, for ratings at a particular age where two or more raters showed no variability across dogs in their ratings, we instead used percent agreement to index reliability. Generally, ICCs are a better method of measuring reliability (see Bartko, 1991) because percent agreement does not differentiate between near agreement and no agreement. Significant differences between inter-observer reliability coefficients (ICC estimates only) for each rating at each age were assessed graphically using means and 95% confidence intervals (Cumming et al., 2007). Initially, we estimated reliability separately from videos collected using the head camera and those collected using the regular camera; however, no significant differences for any rating at any age were found, so we combined ratings across both types of cameras for subsequent agreement analyses.

To evaluate the criterion validity of the non-expert ratings, the non-experts' single and aggregated ratings were correlated with the expert scores. Again we first tested for effects of camera type and found no differences in validity estimates between videos collected using the head camera and those using the regular camera, so we combined results from both types of cameras. For cases in which there was no variability for two or more of the non-expert ratings, we used percent agreement to estimate criterion validity between experts and non-experts. Single estimates of validity were calculated by estimating the correlation between the expert's scores and each individual non-expert's score; the average of each ratings' correlation coefficients was then computed using Fischer's *r*-to-*z* formula. Criterion validity was also examined using aggregate measures, which were calculated by first taking an average of all the non-expert scores and then correlating the aggregate non-expert score with the expert's score. We present both single and aggregate measures in Table 4, but report only single estimates in the text; single measures coefficients are not inflated by the effects of aggregation (Epstein, 1983). Significant differences in criterion validity for each item at each age were assessed graphically using means and 95% confidence intervals (Cumming et al., 2007).

3. Results

3.1. Study 1: inter-observer reliability of MWD ratings

The grand mean of non-expert inter-observer reliability was high for the 10 items (grand mean = 0.96, 95% CIs = 0.91–0.98). Inter-observer reliability across non-experts ranged from 0.89 (vacuum approach) to 0.99 (object possession). Graphical comparisons of confidence intervals showed substantial overlap, indicating that there were no significant differences in overall inter-observer reliability between items (see Online Supplementary Material Fig. 1).

3.2. Study 1: criterion validity

Correlations between expert and non-expert ratings ranged from *r* = 0.57 for object interest to *r* = 0.96 for object possession, with an average across the 10 items of 0.82 (95% CI = 0.62–0.92 for single measures; Table 2). The items with the lowest criterion validity estimates, object interest (*r* = 0.57) and vacuum approach (*r* = 0.64), had confidence intervals that overlapped with less than half of the confidence intervals for the items with the highest validity estimates, object pursuit (*r* = 0.90) and object possession (*r* = 0.96; Fig. 2). There was no difference between any other item.

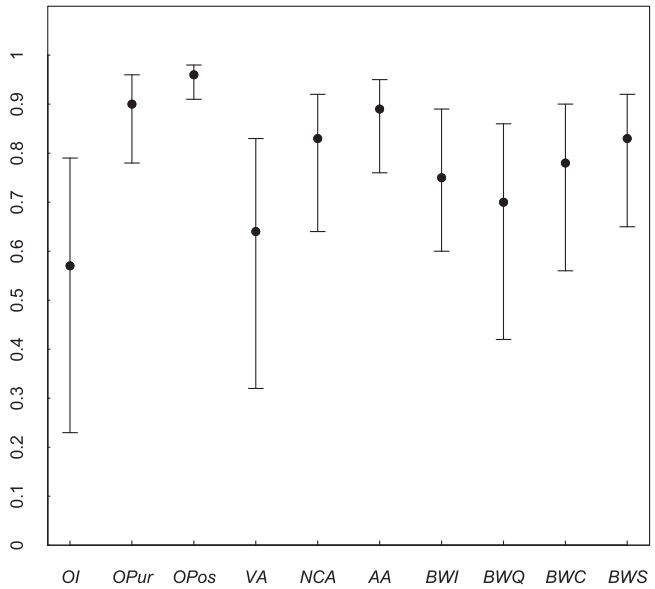

Fig. 2. Single environment correlations and 95% CIs between experts and non-experts to examine criterion validity. OI = object interest, OP = object pursuit, OPos = object possession, VA = vacuum approach, NCA = noisy can approach, AA = avoidance approach, BWI = bitework interest, BWQ = bitework quality, BWC = bitework commitment, BWS = bitework steadiness.

Table 3
Transportation Security Administration Canine Breeding and Development Center behavioral assessment inter-observer reliability estimates and 95% confidence intervals (CIs) using 7non-expert raters in the environment and hunt domain (Study 2).Percent agreement was calculated when two or more raters had no variability in their scores for a particular item and are marked by a percentage. Average intra-class correlation coefficients were calculated by using Fischer's *r* to *z* and then transforming the score back to an *r*. Percent agreements were not used to calculate averages for ages and items, except for in hearing and body sensitivity, where percentages were used for each case, so the average is the average percent agreement.

Item	Domain	3 month (95% CIs)	6 month (95% CIs)	9 month (95% CIs)	12 month (95% CIs)	Item average (95% CIs)
Confidence	Environment	0.91 (0.78–0.97)	0.78 (0.52–0.91)	0.87 (0.71–0.94)	0.90 (0.75–0.96)	0.87 (0.68–0.95)
Concentration	Environment	0.73 (0.41–0.89)	0.67 (0.36–0.85)	0.82 (0.62–0.92)	0.79 (0.52–0.91)	0.76 (0.47–0.90)
Responsiveness	Environment	0.81 (0.56–0.92)	0.88 (0.51–0.90)	0.81 (0.60–0.92)	0.89 (0.72–0.96)	0.85 (0.46–0.90)
Initiative	Environment	0.66 (0.29–0.86)	0.71 (0.42–0.87)	0.79 (0.56–0.91)	0.84(0.62–0.94)	0.76 (0.47–0.90)
Excitability	Environment	0.34 (−0.17–0.71)	72%	0.62 (0.26–0.83)	0.34 (−0.17–0.71)	0.49 (0.04–0.78)
Hearing Sensitivity	Environment	82%	68%	100%	88%	85%
Body Sensitivity	Environment	82%	100%	87%	82%	88%
Chase Retrieve	Environment	0.66 (0.29–0.86)	68%	74%	0.91 (0.77–0.97)	0.82 (0.55–0.93)
Independent Possession	Environment	0.91 (0.78–0.97)	0.72 (0.44–0.87)	0.77 (0.50–0.90)	0.92 (0.80–0.97)	0.85 (0.64–0.94)
Physical Possession	Environment	0.90 (0.75–0.96)	0.73 (0.45–0.87)	0.85 (0.67–0.93)	0.86 (0.67–0.95)	0.84 (0.62–0.94)
Age average	Environment	0.79 (0.50–0.92)	0.76 (0.48–0.90)	0.80 (0.56–0.91)	0.85 (0.65–0.94)	0.80 (0.54–0.92)
Chase Retrieve	Hunt	0.67 (0.40–0.83)	0.51 (0.16–0.75)	0.49 (0.11–0.74)	−0.24 (−0.62–0.23)	0.39 (−0.03–0.69)
Mental Possession	Hunt	0.72 (0.47–0.86)	0.62 (0.31–0.81)	0.67 (0.37–0.84)	0.62 (0.24–0.83)	0.66 (0.34–0.84)
Physical Possession	Hunt	0.74 (0.48–0.88)	0.92 (0.83–0.96)	0.94 (0.87–0.97)	0.95 (0.87–0.98)	0.91 (0.79–0.96)
Hidden Grass	Hunt	0.81 (0.63–0.91)	0.86 (0.71–0.93)	0.90 (0.83–0.95)	0.81 (0.57–0.92)	0.85 (0.67–0.93)
Independent Possession	Hunt	0.76 (0.54–0.88)	0.76 (0.53–0.88)	0.94 (0.87–0.97)	0.86 (0.67–0.94)	0.85 (0.67–0.93)
Hidden 1	Hunt	0.98 (0.96–0.99)	0.95 (0.89–0.98)	0.98 (0.95–0.99)	0.33 (−0.13–0.67)	0.93 (0.84–0.97)
Hidden 2	Hunt	0.96 (0.91–0.98)	0.97 (0.93–0.99)	0.98 (0.96–0.99)	0.90 (0.76–0.96)	0.96 (0.91–0.98)
Activity	Hunt	0.66 (0.37–0.83)	0.49 (0.14–0.73)	0.65 (0.34–0.83)	−0.24 (−0.62–0.23)	0.43 (0.02–0.72)
Age average	Hunt	0.81 (0.61–0.91)	0.83 (0.65–0.92)	0.90 (0.78–0.96)	0.66 (0.28–0.86)	0.82 (0.61–0.92)

3.3. Study 2: inter-observer reliability of TSA-CBDC dog ratings

The grand mean of non-expert inter-observer reliabilities was high for the 10 items in the environment assessments (grand mean = 0.80, 95% CIs = 0.54–0.92) and the eight items in the hunt assessments (grand mean = 0.82, 95% CIs = 0.61–0.92) (Table 3). In the environment assessment, inter-observer reliability across non-experts ranged from 0.49 (excitability) to 0.87 (confidence) and there were no significant differences between the items using inter-observer reliability collapsed across ages, $p > 0.05$ based on Cumming et al. (2007) (see Online Supplementary Material Fig. 2). Across items at each age, the average inter-observer reliability was 0.79 at 3-months, 0.76 at 6-months, 0.80 at 9-months, and 0.85 at 12-months (averages were computed using only ratings that allowed computation of an ICC, and not ratings which required percent agreement estimates). Graphical comparisons of confidence intervals indicated that there were no significant differences in overall inter-observer reliability between ages, $p > 0.05$ (see Online Supplementary Material Fig. 3).

For the hunt assessment ratings collapsed across ages, inter-observer reliability ranged from 0.39 (chase retrieve) to 0.96 (hidden two) and there were significant differences in inter-observer reliability between the items with the lowest inter-observer reliability (chase retrieve, mental possession, and activity) and the items with the highest inter-observer reliability (hidden 1 and hidden 2) (see Online Supplementary Material Fig. 4). For estimates collapsed across ratings within age groups, average inter-observer estimates were 0.81 at 3-months, 0.83 at 6-months, 0.90 at 9-months, and 0.66 at 12-months. Graphical comparisons of confidence intervals indicated that there were no significant differences in inter-observer reliability between ages, $p > 0.05$ (see Online Supplementary Material Fig. 5).

3.4. Study 2: criterion validity

Correlations between expert and non-expert ratings were 0.46 (95% CIs = 0.00–0.76) for the environment assessment domain and 0.59 (95% CIs = 0.23–0.81) for the hunt assessment domain (Table 4). In the environment assessment, criterion validity estimates for single ratings ranged from $r = 0.19$ for excitability to $r = 0.77$ for chase retrieve. The item with the lowest validity estimate ($r = 0.19$; excitability) had a confidence interval that overlapped with less than half of the confidence interval for the item with the highest validity estimate ($r = 0.77$; chase retrieve; see Online Supplementary Material Fig. 6).

In the hunt assessment, criterion validity estimates for single ratings between experts and non-experts ranged from $r = 0.20$ for mental possession to $r = 0.89$ for hidden 1. The items with the lowest validity (mental possession and chase retrieve) had confidence intervals that overlapped with less than half of the confidence intervals for the items with the highest validity estimates (hidden 1 and hidden 2; see Online Supplementary Material Fig. 7). Across items at each age, the average environment validity estimates were 0.46 for 3-months, 0.35 for 6-months, 0.45 for 9-months, and 0.55 for 12-months (Fig. 3). Across items at each age, the average hunt validity estimates were 0.58 for 3-months, 0.55 for 6-months, 0.65 for 9-months, and 0.56 for 12-months (Fig. 4). In both domains, graphical comparisons of confidence intervals indicated that there were no significant differences in criterion validity between ages, $p > 0.05$ (see Online Supplementary Materials Figs. 8 and 9).

4. Discussion

In two independent studies, we evaluated whether minimally trained non-experts behavioral ratings were reliable and valid in working-dog populations. We tested reliability by examining inter-observer reliability across non-expert raters. We tested validity by examining criterion validity between experts and non-experts. Thus, validity in this case indexes only whether non-experts match expert ratings and does not provide any data on whether the non-expert or expert ratings reflect real attributes of the dogs. The reliabilities were generally good in both studies, and many items exceeded the 0.70 threshold sometimes proposed as a threshold for acceptability. The validities were also relatively strong. Our results suggest that non-experts are capable of reliably rating dog behavior and that their ratings can reproduce expert ratings moderately well, with moderate variability in reliability and validity across items. Even for items that we found to be less reliable, the findings provide the necessary information to determine which items may need more than one rater to be rated reliably. Specifically, using the Spearman–Brown prophesy formula, researchers can use

Table 4
Transportation Security Administration Canine Breeding and Development Center environment and huntassessment single estimate correlations between expert and non-expert scores. Aggregate estimates are listed in parentheses. Percent agreement was calculated when two or more raters had no variability in their scores for a particular item. Averages were calculated by using Fischer's *r* to *z* and then transforming the score back to an *r*. Percent agreements were not used in calculating averages, except for in hearing body sensitivity, where percentages were used for each case, so the average is the average percent agreement.

Item	Domain	3 month	6 month	9 month	12 month	Item average
Confidence	Environment	0.28 (0.34)	0.37 (0.52)	0.47 (0.61)	0.45 (0.54)	0.40 (0.51)
Concentration	Environment	0.20 (0.24)	0.10 (0.15)	0.37 (0.53)	0.54 (0.75)	0.31 (0.45)
Responsiveness	Environment	0.41 (0.56)	0.65 (0.77)	0.58 (0.81)	0.58 (0.72)	0.56 (0.73)
Initiative	Environment	0.37 (0.63)	0.19 (0.26)	0.53 (0.80)	0.47 (0.59)	0.40 (0.60)
Excitability	Environment	0.13 (0.14)	68%	0.27 (0.30)	0.13 (0.14)	0.19 (0.19)
Hearing Sensitivity	Environment	94%	68%	96%	82%	85%
Body Sensitivity	Environment	76%	100%	83%	82%	85%
Chase Retrieve	Environment	0.60 (0.75)	64%	74%	0.87 (0.99)	0.77 (0.82)
Independent Possession	Environment	0.80 (0.88)	0.34 (0.54)	0.44 (0.65)	0.59 (0.70)	0.57 (0.720
Physical Possession	Environment	0.63 (0.76)	0.37 (0.51)	0.48 (0.68)	0.51 (0.67)	0.50 (0.66)
Age average	Environment	0.46 (0.59)	0.35 (0.49)	0.45 (0.65)	0.55 (0.74)	0.46 (0.65)
Chase Retrieve	Hunt	0.49 (0.68)	−0.07 (−0.15)	0.17 (0.19)	0.28 (0.66)	0.28 (0.39)
Mental Possession	Hunt	−0.03 (−0.10)	0.22 (0.34)	0.43 (0.72)	0.17 (0.32)	0.20 (0.36)
Physical Possession	Hunt	0.51 (0.75)	0.68 (0.80)	0.72 (0.82)	0.76 (0.84)	0.76 (0.84)
Hidden Grass	Hunt	0.26 (0.35)	0.58 (0.76)	0.75 (0.92)	0.45 (0.58)	0.53 (0.72)
Independent Possession	Hunt	46%	0.47 (0.70)	0.76 (0.87)	0.53 (0.69)	0.60 (0.77)
Hidden 1	Hunt	0.90 (0.95)	0.85 (0.92)	0.93 (0.97)	72%	0.89 (0.95)
Hidden 2	Hunt	0.84 (0.91)	0.88 (0.93)	0.45 (0.48)	0.85 (0.93)	0.79 (0.88)
Activity	Hunt	0.55 (0.54)	0.15 (0.23)	0.48 (0.69)	67%	0.41 (0.51)
Age average	Hunt	0.58 (0.69)	0.55 (0.68)	0.65 (0.79)	0.56 (0.72)	0.59 (0.72)

the current findings to determine how many non-expert raters would be needed to attain any desired reliability threshold (Block, 1961). With this information, practitioners and researchers can then evaluate the relative viability of using expert vs. non-expert raters given their own practical and financial parameters. Thus, the findings have significant practical implications for working-dog programs currently using experts to evaluate behavior in dogs.

There were a number of key differences between Studies 1 and 2. The two studies differed in terms of breeds studied (with Study 1 focusing using Belgian Malinois and Study 2 using Labrador Retrievers, Viszlas, and Labrador Retriever/German Shorthaired Pointers), assessment conditions (Study 1 undertook the assessments in a much more controlled environment than those used in Study 2), and age groups studies (in Study 1 dogs were examined at 2-months of age but in Study 2 dogs were examined at 3, 6, 9, and 12-months). The fact that we found relatively good reliability and validity in both studies underscores the robustness of the effects.

4.1. Potential threats to reliability and validity

Despite the broad consistencies in the findings, some items were assessed with lower levels of reliability and validity than were others. In Study 1, we found vacuum approach and object interest were characterized by lower validity than other items. In Study 2, we found chase retrieve, mental possession, and activity were characterized by lower reliability than other items and excitability, mental possession, and chase retrieve were characterized by lower validity than other items.

What could account for the lower reliability and validity associated with some items? One possibility is suggested from previous research, which has shown that dog experts gaze at the dog's body more than non-dog experts do; non-experts gaze mostly at the dog's head (Kujala et al., 2012). Such differences in human gaze focus could come into play when interpreting dogs' responses to fearful stimuli (e.g., in vacuum approach), if items associated with fear are communicated by dog postural changes and body language (Coren, 2000). However, it should be noted that our results indicating that vacuum approach was difficult to measure accurately are not consistent with previous research, which suggest fearfulness is one of the more easily recognized behaviors in dogs (Bahlig-Pieren and Turner, 1999; Tami and Gallagher, 2009).

Variation in the findings may have also occurred because of low visibility of some items. For example, in humans some items (e.g., talkativeness) are much more visible and can be seen across more situations than other items (e.g., creativity). Visibility has been shown to be associated with the reliability with which an item is judged (Funder and Dobroth, 1987). In this sense, video recordings may have rendered some items less visible than others. For example, in the hunt assessment the camera may not have captured the dog's full focus on the towel, which was needed to rate the item mental possession. Low visibility could have caused non-experts to disagree on how to rate the item. Further, experts were able to view the dog's focus on the towel in person, so they were likely able to have a better view of the dog's mental possession. These differences could have led to a discrepancy between how experts and non-experts rated mental possession. These results highlight the importance of recording videos that capture the full range of an item in order to ensure good reliability and validity.

Similar to visibility, another reason for lower criterion validity in some items may be due to the way in which experts vs. non-experts observe behavior. Not surprisingly, experts are often better than non-experts in discriminating between levels of behavior that are subtle (e.g., in observing differences in physical conditions; Kristensen et al., 2006). In Study 1, differences between levels of ratings may have been too subtle for non-experts to discriminate based on the written description of some items. For example, differences between a '3' and a '4' for the item 'object interest' were relatively subtle, with the words changing from 'looks at object with some intensity' for a '3' to 'looks at object and follows its movements' for a '4'. The subtle differences may have made it difficult for non-experts to make distinctions and rate dogs according to the same standards that experts were using. Further, it is likely there were subtle differences between levels of ratings for items in Study 2 that had little variability overall that non-experts were unable to detect. For example, excitability was rated based on average excitability within the TSA dog population. In this case, it is not surprising that experts for this study were more likely to notice subtle differences in excitability compared to non-experts. Thus, some items may need to either be rated only be experts or non-experts must be taught to distinguish between subtle differences in behavior to obtain the highest validity of ratings.

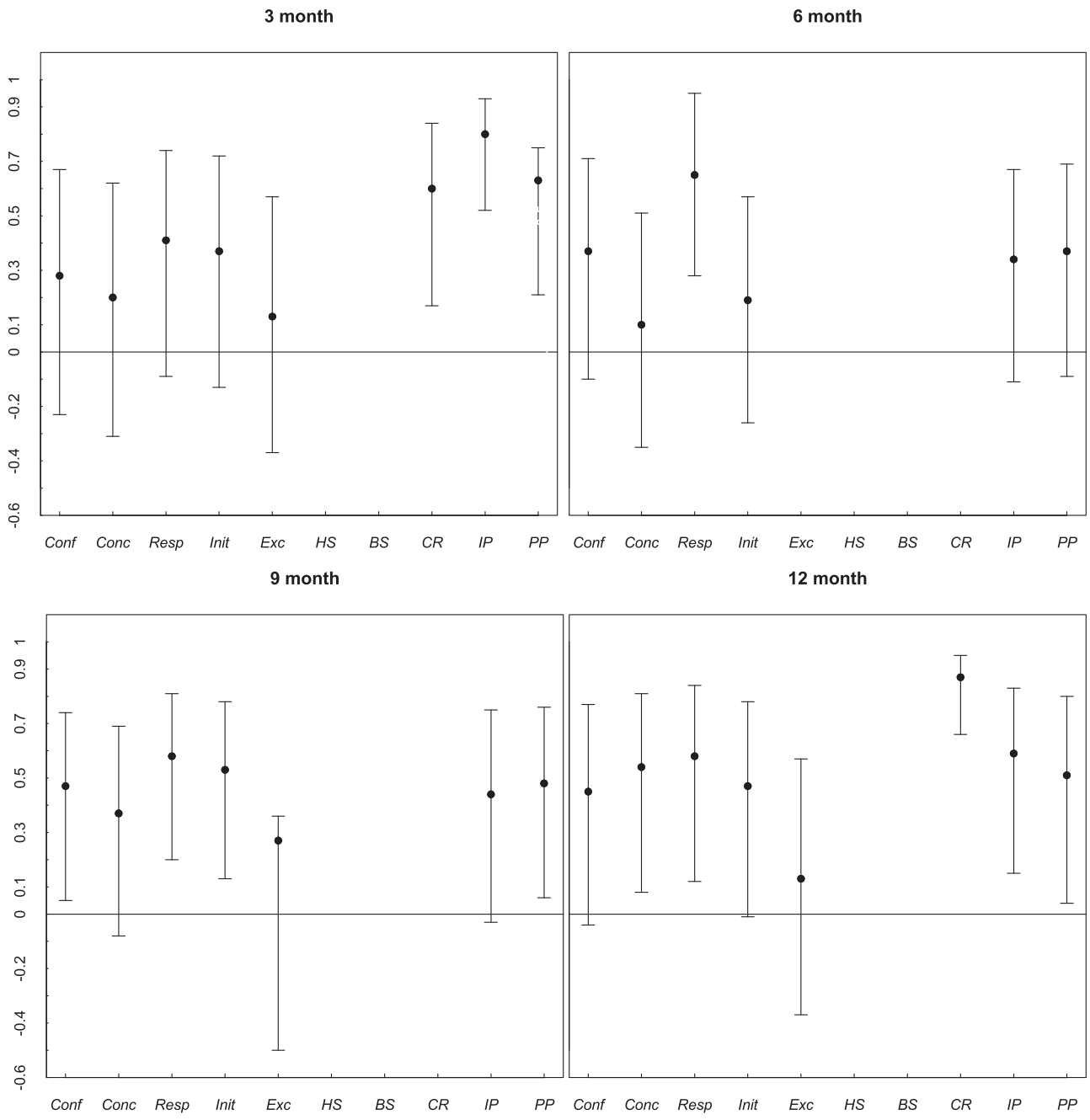

Fig. 3. Single environment correlations and 95% CIs between experts and non-experts to examine criterion validity at 4 different ages. Items with no mean or CIs were calculated using percent agreement, so have no means and CIs plotted in these graphs. Conf = confidence, Conc = concentration, Resp = responsiveness, Init = initiative, Exc = excitability, HS = hearing sensitivity, BS = body sensitivity, CR = chase retrieve, IP = independent possession, PP = physical possession.

Another possible reason for why some items may have shown low reliability may be because some behaviors may just not be reliable in these assessments. Due to a limited number of true experts in these organizations, we did not examine inter or intra-observer reliability between experts. For items in which reliability was low among non-experts, experts may also have had trouble reliably rating those items in the assessments. Previous research provides conflicting evidence, with some research suggesting experienced animal welfare inspectors show good inter-observer reliability when observing sheep behavior (Phythian et al., 2013) and other research suggesting novice raters show higher inter-observer reliability than expert raters for all items when observing pig behavior (Cloudard et al., 2011). We did find relatively good reliability for most items in both studies, but some behaviors may not have been reliable in these assessments.

One of the limitations of the present work was the lack of variability in the ratings for many of the items observed in Study 2, which could have influenced the reliability and validity differences in items. For example, the expert gave '5's to 23 out of 26 of the 3-month-old dogs for the item of 'activity'. In this case, most, if not all of the dogs in the TSA-CBDC population could be described as having very high activity levels during assessments. In humans, people with higher levels of behavioral activity should be easier to judge than people who have less activity because more activity affords more clues about their personality (Funder, 1995). In our study, non-experts did not have much of an opportunity to observe

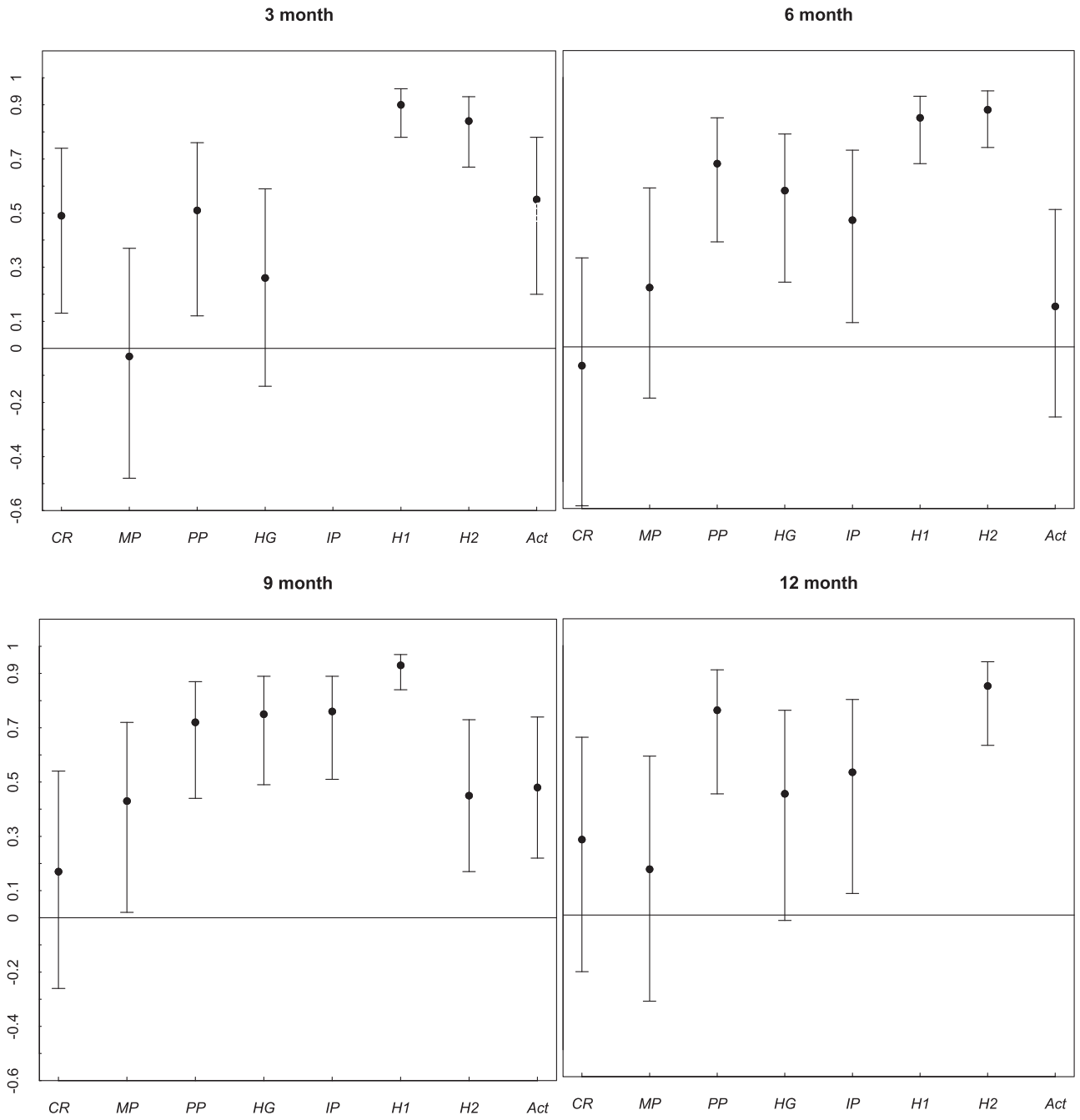

Fig. 4. Single hunt correlations and 95% CIs between experts and non-experts to examine criterion validity at 4 different ages. Items with no mean or CIs were calculated using percent agreement, so have no means and CIs plotted in these graphs. CR = chase retrieve, MP = mental possession, PP = physical possession, HG = hidden grass, IP = independent possession, H1 = hidden 1, H2 = hidden 2, ACT–activity.

a dog truly low on several items, which may have resulted in raters searching for differences in how a dog performed even when those differences were not present. Low variability is related to low inter-observer reliability due to restriction of range issues (Allik et al., 2010), so variability in behavior is essential to reliably rate an item. Chase retrieve, excitability, and activity were all characterized by low variability, which may have resulted in both lower reliability and lower validity in this study. Interestingly, a lack of variation in some items may also be an aspect of successful breeding in a working-dog population, and not necessarily viewed as a negative property of a working-dog population.

Another limitation is that the non-experts in this study were not complete non-experts because they did receive some training. To determine whether true novices can reliably and validly rate assessments, future research should compare the levels of reliability and validity obtained from observers with different levels of training, including a condition with no training at all. Such research is needed to determine the optimal tradeoff between investments in training and decreases in reliability and validity.

We had predicted that reliability and validity might vary across age because consistency varies across age (Fratkin et al., 2013). However, we found no significant differences in reliability and validity based on the age of the dog. Little research has examined age-based difference in inter-observer reliability and validity of behavioral ratings of dogs. Our results suggest behavioral ratings of dogs can show good reliability and validity when

non-experts rate the assessments, regardless of the age of the dog.

5. Conclusion

Our results suggest minimally trained non-experts can reliability rate dog behavior from behavioral assessments and non-experts can reproduce expert ratings of behavior in dogs at several different ages. However, despite the generally promising results, some items were rated with less reliability and validity. Potential factors for improving the reliability and validity with which such items are rated include making sure items are captured completely via video recordings, making sure non-experts are trained thoroughly to detect items that are subtle to differentiate, and making sure there is some variability in the behavior of the subjects being assessed. If these steps are followed, the present research can lay the groundwork for using non-experts to undertake behavioral assessments of working dogs; such non-expert assessments could be crucial in many applied contexts because they could significantly reduce the burden on the expert assessors.

Acknowledgments

Funding was provided by the US Department of Homeland Security, Contact HSHQDC-10-C-00085 and the National Science Foundation Award 0731216. We thank Christy Amador, Jordan Buckley, Justin Chumbley, Kristen Cunningham, Dominique Egger, Rebekah Ellis, Jade Fountain, Joshua Goldberg, Mario Guerra, Hayden Henderson, Charlotte Horne, Sanaa Karim, Sarah Kerwin, Ilissa Madrigal, Lauren Miller, Taylor Mezaraups, Aamna Najam, Dorothee Rocznik, Sarah Skidmore, and Carishia Williams for their help in rating videos for this study. In addition, we thank Miles Bensky, Stephen DeBono, Monica McGarrity, and Diana Thomas for their input on ideas for the paper and help with the collection of data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.beproc.2014.09.028.

References

Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., Metspalu, A., 2010. Variance determines self-observer agreement on the Big Five personality traits. J. Res. Pers. 44, 421–426.
Bahlig-Pieren, Z., Turner, D.C., 1999. Anthropomorphic interpretations and ethological descriptions of dog and cat behavior by lay people. Anthrozoos 12, 205–210.
Bartko, J.J., 1991. Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17, 483–489.
Block, J., 1961. The Q-sort Method in Personality Assessment and Psychiatric Research. Thomas, Springfield, IL.
Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.
Cloudard, C., Meunier- Salaün, M., Devillers, N., 2011. Development of approach and handling tests for the assessment of reactivity to humans of sows housed in stall or in group. Appl. Anim. Behav. Sci. 133, 26–39.
Coren, S., 2000. How to Speak Dog. Free Press, New York.
Cumming, G., Fidler, F., Vaux, D.L., 2007. Error bars in experimental biology. J. Cell. Biol. 177, 7–11.
Duncan, L.M., Pillay, N., 2012. Volunteer experience influences the conclusions of behavioral experiments. Appl. Anim. Behav. Sci. 140, 179–187.
Epstein, S., 1983. Aggregation and beyond: some basic issues on the prediction of behavior. J. Pers. 51, 360–392.
Fratkin, J.L., Sinn, D.L., Patall, E.A., Gosling, S.D., 2013. Personality consistency in dogs: a meta-analysis. PLOS ONE 8 (1), e54907.
Funder, D.C., 1995. On the accuracy of personality judgment: a realistic approach. Psychol. Rev. 102, 652–670.
Funder, D.C., Dobroth, K.M., 1987. Differences between traits: properties associated with interjudge agreement. J. Pers. Soc. Psychol. 52 (2), 409–418.
Gosling, S.D., 2001. From mice to men: what can we learn about personality from animal research? Psychol. Bull. 127, 45–86.
Gosling, S.D., John, O.P., Craik, K.H., Robins, R.W., 1998. Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. J. Pers. Soc. Psychol. 74, 1337–1349.
John, O.P., Robins, R.W., 1993. Determinants of interjudge agreement on personality traits: the big five domains, observability, evaluativeness, and the unique perspective on the self. J. Pers. 61, 521–551.
John, O.P., Soto, C.J., 2007. The importance of being valid. In: Robins, R.W., Fraley, R.C., Krueger, R. (Eds.), Handbook of Research Methods in Personality Psychology. The Guilford Press, New York, pp. 461–494.
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.
Kristensen, E., Dueholm, L., Vink, D., Anderson, J.E., Jakobsen, E.B., Illum-Nielsen, S., Peterson, F.A., Enevoldsen, C., 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. J. Dairy Sci. 89, 3721–3728.
Kujala, M.V., Kujalam, J., Carlson, S., Hari, R., 2012. Dog experts' brains distinguish socially relevant body postures similarly in dogs and humans. PLoS ONE 7, e39145.
Maejima, M., Inoue-Murayama, M., Tonosaki, K., Matsuura, N., Kato, S., Saito, Y., Weiss, A., Murayama, Y., Ito, S., 2007. Traits and genotypes may predict the successful training of drug detection dogs. Appl. Anim. Behav. Sci. 107, 287–298.
Martau, P.A., Caine, N.G., Candland, D.K., 1985. Reliability of the Emotions Profile Index, primate form, with *Papio hamadryas*, *Macaca fuscata*, and two Saimiri species. Primates 26, 501–505.
Mirkó, E., Dóka, A., Miklósi, A., 2013. Association between subjective rating and behaviour coding and the role of experience in making video assessments on the personality of the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci. 149, 45–54.
Paroz, C., Gebhardt-Henrich, S.G., Steiger, A., 2008. Reliability and validity of behaviour tests in Hovawart dogs. Appl. Anim. Behav. Sci. 115, 67–81.
Phythian, C., Michalopoulou, E., Duncan, J., Wemelsfelder, F., 2013. Inter-observer reliability of qualitative behavioural assessments of sheep. Appl. Anim. Behav. Sci. 144, 73–79.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 2, 420–428.
Sinn, D.L., Gosling, S.D., Hilliard, S., 2010. Personality and performance in military working dogs: reliability and predictive validity of behavioral tests. Appl. Anim. Behav. Sci. 127, 51–65.
Sinn, D.L., Gosling, S.D., Moltschaniwskyj, N.A., 2008. Development of shy/bold behaviour in squid: context-specific phenotypes associated with developmental plasticity. Anim. Behav. 75, 433–442.
Tami, G., Gallagher, A., 2009. Description of the behaviour of domestic dog (*Canis familiaris*) by experienced and inexperienced people. Appl. Anim. Behav. Sci. 120, 159–169.
Waller, B.W., Dunbar, R.I.M., 2005. Differential behavioural effects of silent bared teeth display and relaxed open mouth display in chimpanzees (*Pan troglodytes*). Ethology 111, 129–142.
Wan, M., Bolger, N., Champagne, F.A., 2012. Human perception of fear in dogs varies according to experience with dogs. PLoS ONE 7 (12), e51775.
Wemelsfelder, F., Hunter, A.E., Paul, E.S., Lawrence, B., 2012. Assessing pig body language: agreement and consistency between pig farmers, veterinarians, and animal activists. J. Anim. Sci. 90, 3652–3665.