Applied Animal Behaviour Science 127 (2010) 51–65

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

Personality and performance in military working dogs: Reliability and predictive validity of behavioral tests

David L. Sinn[a,c], Samuel D. Gosling[a,*], Stewart Hilliard[b]

[a] University of Texas, Department of Psychology, 1 University Station A8000, Austin, TX 78712-0187, USA
[b] United States Air Force, 341st Training Squadron/Logistics, 1320 Truemper St., Ste 2, Lackland AFB, TX 78236-5502, USA
[c] University of Tasmania, School of Zoology, Private Bag 5, Hobart, TAS, 7001, Australia

ARTICLE INFO

Article history:
Accepted 13 August 2010
Available online 9 September 2010

Keywords:
Military dog
Personality
Reliability
Predictive validity
Behavioral instrument

ABSTRACT

Quantification and description of individual differences in behavior, or personality differences, is now well-established in the working dog literature. What is less well-known is the predictive relationship between particular dog behavioral traits (if any) and important working outcomes. Here we evaluate the validity of a dog behavioral test instrument given to military working dogs (MWDs) from the 341st Training Squadron, USA Department of Defense (DoD); the test instrument has been used historically to select dogs to be trained for deployment. A 15-item instrument was applied on three separate occasions prior to training in patrol and detection tasks, after which dogs were given patrol-only, detection-only, or dual-certification status. On average, inter-rater reliability for all 15 items was high (mean = 0.77), but within this overall pattern, some behavioral items showed lower inter-rater reliability at some time points (<0.40). Test–retest reliability for most (but not all) single item behaviors was strong (>0.50) across shorter test intervals, but decreased with increasing test interval (<0.40). Principal components analysis revealed four underlying dimensions that summarized test behavior, termed here 'object focus', 'sharpness', 'human focus', and 'search focus'. These four aggregate behavioral traits also had the same pattern of short-, but not long-term test–retest reliability as that observed for single item behaviors. Prediction of certification outcomes using an independent test data set revealed that certification outcomes could not be predicted by breed, sex, or early test behaviors. However, prediction was improved by models that included two aggregate behavioral trait scores and three single item behaviors measured at the final test period, with 1 unit increases in these scores resulting in 1.7–2.8 increased odds of successful dual- and patrol-only certification outcomes. No improvements to odor-detection certification outcomes were made by any model. While only modest model improvements in prediction error were made by using behavioral parameters (2–7%), model predictions were based on data from dogs that had successfully completed all three test periods only, and therefore did not include data from dogs that were rejected during testing or training due to behavioral or medical reasons. Thus, future improvements to predictive models may be more substantial using independent predictors with less restrictions in range. Reports of the reliability and validity estimates of behavioral instruments currently used to select MWDs are scarce, and we discuss these results in terms of improving the efficiency by which working dog programs may select dogs for patrol and odor-detection duties using behavioral pre-screening instruments.

© 2010 Elsevier B.V. All rights reserved.

* Corresponding author. Tel.: +1 5124711628; fax: +1 5124715935.
E-mail address: samg@mail.utexas.edu (S.D. Gosling).

0168-1591/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.applanim.2010.08.007

1. Introduction

1.1. Behavioral variation and reliability and predictive validity of measurement instruments

Military working dogs (MWDs) are used by numerous law-enforcement and governmental organizations for a variety of purposes, including police/patrol duties, and explosive and narcotics detection. In the current worldwide political climate, there is an increasing need for specialist working dogs, especially those that can be used for personnel detection (i.e., search a building for intruders) and apprehension, and detection of dangerous substances, such as poisons, gunpowder, and other explosives. Not surprisingly, the science of working dogs is currently an active area of research (Helton, 2009).

One recurring observation from this work is that some working dogs are better at their assigned tasks than are others, and these differences in performance are due mainly to behavioral rather than sensory or morphological differences (Slabbert and Odendaal, 1999; Svartberg, 2002; Maejima et al., 2007; Rooney et al., 2007). In many selection and training programs for police and detection dogs, more than half of the candidate dogs are rejected for behavioral reasons (Wilsson and Sundgren, 1997b; Slabbert and Odendaal, 1999; Maejima et al., 2007). Given the high costs of selecting, training, and deploying specialist working dogs, the outstanding issue facing agencies interested in training working dogs is quantifying variation in behavior empirically and understanding how differences in behaviors predict subsequent performance in working environments.

Several methods for quantifying dog behavior are currently in use. Some methods are based on ratings of observed behaviors, such as observers scoring the presence/absence of particular postures or biting to quantify aggression (e.g., Haverbeke et al., 2009). More commonly, however, observers familiar with subject animals assign subjective ratings to individual dogs based on their performance in standardized behavioral tests (e.g., Svartberg, 2005; Ley et al., 2008). For example, observers can use a scale from one to six to rate the 'confidence' of dogs when encountering a novel intruder, with lower scores indicating dogs that behave in a hesitant or fearful manner, and higher scores indicating bold, confident behavior. Almost all studies on working dogs use a large number of behavioral measures to attempt to capture an individual's relevant overall behavioral style, or its personality (reviewed in Jones and Gosling, 2005). For example, instead of evaluating single behavioral responses in single tests, most working dog programs attempt to quantify large numbers of measured behaviors which are combined into a smaller number of aggregate traits that meaningfully summarize an individual's disposition, usually through data reduction techniques such as principal components analysis (PCA, Svartberg and Forkman, 2002; Svartberg et al., 2005; Ley et al., 2008).

In order for behavioral measurement instruments to be valuable, however, they must be both reliable and valid (Gosling et al., 2003; Svartberg, 2005). Two core forms of reliability are inter-rater agreement and test–retest reliability. Inter-rater reliability is an index of the extent to which different observers describe the same individual the same way. Surprisingly, reports of inter-rater reliability in working dogs are scarce (see Graham and Gosling, 2009), and the few published reports provide mixed evidence. In one study of companion dogs (Ley et al., 2009), inter-observer reliabilities on five aggregated behavioral traits (measured by the 26-item MCPQ-R scale) were high, with an overall average of 0.62. In contrast, inter-rater agreement on 14 single behavioral items in guide dogs ranged from 0.00 for 'willingness to carry out commands' to 0.7 for 'nervousness of people, traffic, and strange places' (Goddard and Beilharz, 1983). Thus, while inter-rater reliability of behavioral measurement instruments in many non-human animals, including companion dogs, can be good (0.52–0.62: Gosling, 2001; Gosling et al., 2003), it cannot be assumed to be strong. In fact, we are not aware of any assessments of inter-rater reliability for measurement instruments used to select and certify patrol and detection dogs.

Test–retest reliability describes the extent to which individuals' scores generalize across testing occasions. Reports of test–retest reliability are scarce for patrol and detection dogs, but reports from guide and companion dogs are mixed, and suggest that test–retest reliability may be trait- or study-specific. In some companion dogs, high test–retest correlations (>0.60) have been reported for traits such as extraversion, sociability, aggressiveness, neuroticism, and curiosity/fearlessness over test periods from 1 to 6 months (Netto and Planta, 1997; Svartberg et al., 2005; Ley et al., 2009), but in other studies of guide dogs tested over a 6 month period test–retest correlations for distractibility and activity were lower (<0.40), and for aggression were negligible (0.20: Goddard and Beilharz, 1984; Goddard and Beilharz, 1985).

A third criterion – ultimately the most important – for the usefulness of behavioral measurement instruments for working dog programs is predictive validity. Behavioral instruments used at one time (such as for selection of dogs for training) should predict certification or work-related outcomes at another time if they are to be useful for working dog agencies. Not surprisingly, the ability of a battery of behavioral measures taken at one time to predict working outcomes at another has received the greatest attention in working-dog studies. Perhaps the most well-known examples are those involving 'puppy tests', where prospective guide or shelter puppies are given a battery of tests designed to measure traits such as fearfulness, sociability, distractedness, and aggression (Svartberg, 2005), and these test results are then compared to training certification outcomes (in the case of guide dogs, Goddard and Beilharz, 1983, 1986) or problem behaviors in the home (in the case of companion dogs, van der Borg et al., 1991; Hennessy et al., 2001). Results of the predictability of later behaviors measured by earlier behavioral instruments in patrol and detection dogs are mixed. In some police dog training programs, the predictability of adult behavior from earlier puppy behaviors was low, and not different from chance (Wilsson and Sundgren, 1998). In other programs, high percentages of successful (92%) and unsuccessful (82%) certified dogs could be characterized by three out of eight

puppy personality tests (Slabbert and Odendaal, 1999). In drug-detection dogs, high scores for the trait 'desire for work' characterized 93% of dogs that passed training but only 53% of dogs that did not pass (the other aggregate trait measured in this study, 'distractibility', did not predict training certification, Maejima et al., 2007). It is still unclear which behavioral traits in patrol and detection dogs result in positive certification and working life outcomes.

To date, no studies of patrol or detection dogs have provided estimates of inter-rater and test–retest reliability, and predictive validity (using independent test data) of their measurement instruments. Here, we report such findings.

1.2. Current sample and aims

The 341st Training Squadron at Lackland Air Force Base, Texas is the USA Department of Defense Executive Agent for providing trained military working dogs and handlers to the United States Air Force, Army, Marines, and Navy. MWDs are assigned to each branch of service's security police, who provide "force protection" for service assets and personnel. MWD procurement in this program is based on a behavioral test instrument. Behavioral tests used in the Lackland MWD program, much like tests used in other working dog programs, are based on traditional methods used by lay dog trainers since the early 1900s to select dogs for breeding (e.g., Ruefenacht et al., 2002; Fuchs et al., 2005). These tests are meant to assess the strength of a dog's motivation to engage in goal-directed behaviors using positive reinforcement. If dogs are chosen for procurement, they undergo training and certification tests, after which a given MWD may be dual-certified, detection certified but patrol eliminated, detection eliminated but patrol certified, or dual-eliminated. Of 660 MWDs that finished training at Lackland between October 2007 and September 2009, 63.5% certified dual, 15.6% certified in detection or patrol, and 20.9% were dual-eliminated. In other words, fully 36.5% of this two-year sample failed either one or both phases of training. Given the individual cost of dog procurement and training, there is a need for the Lackland MWD program to efficiently procure, train, and select dogs that are likely to be certified and to succeed in subsequent field working conditions.

Specifically, the aims of the current study were to: (1) evaluate the inter-rater and test–retest reliability of behaviors used in the current Lackland MWD behavioral measurement instrument, (2) evaluate the usefulness of aggregate PCA scores as a representation of MWD behavior, and (3) predict the certification outcomes of MWDs from earlier test behavior using two independent data sets, the first to generate model parameters, and the second to generate estimates of prediction error.

2. Methods

2.1. Subjects

Subjects were procured from vendors in Europe on multiple buy trips from March 2006 to August 2008. Most were German Shepherd Dogs (GS; $N = 735$) and Belgian Sheepherd Dogs, of the variety Malinois (BM; $N = 243$), but some Dutch Shepherd (DS; $N = 22$) individuals were also procured. There were 125 female and 610 male GS subjects, 85 female and 158 male BM subjects, and eight female and 14 male DS dogs. GSDs were procured from nine different vendor sites, BM from 14 different vendor sites, and DS from six different vendor sites, but 96%, 90%, and 95% of individuals of each breed, respectively, were procured from the same five vendors. Specific birthdates of animals were not available from the vendor but animals were typically 1 to 3 years of age at the time of purchase. Once purchased from the vendor site, animals were consigned to a specialized transport company, and were transported in large plastic airline crates according to airline standards to Lackland Air Force Base, San Antonio, Texas. Dogs were then trained for approximately 100 days (mean = 108 days, SD = 34) before being deemed ready for two certification tests, one each for patrol and odor-detection duties. Patrol certification consisted of performance testing in agility (jumping, negotiating catwalks, tunnels, walls, and stairways), obedience (sitting, lying down, heeling, and remaining steady under blank gunfire), controlled aggression (guarding, pursuing and apprehending, releasing on command, escorting and protecting the handler), and searching (searching out human decoys). Odor-detection certification consisted of a series of searches in different operational areas (vehicles, barracks, theater, aircraft, etc.). Odor-detection certification was tested over 3 days, during which dogs were required to correctly indicate at least 19 of 20 different explosive "training aids," while exhibiting no more than two false responses.

2.2. General behavioral test procedure

Standardized behavioral tests, based on pre-existing procurement tests and previous studies, have been in use by the Lackland MWD program since 1992 (Supplementary data, Appendix A, Champness, 1996; Wilsson and Sundgren, 1997b). It consists of 15 individual item ratings designed to assess subject dogs' proficiency in four working domains. Two working domains, environmental and gun sureness, consist of single items, while the other two domains, detection and patrol, consist of six (static object, thrown object, possession, physical possession, search activity, and search stamina) and seven (defense, threat aggression, non-threat bite quality, threat bite quality, attention transfer, pursuit bite, and frontal bite) individual items, respectively. Nine hundred seventy-seven dogs were assessed using this test instrument at the vendor site in Europe previous to procurement (hereafter termed 'test 1'), a sub-set of the dogs from test 1 were assessed at an off-vendor site in Europe prior to transport to the USA (65 dogs; hereafter termed 'test 2'), while 477 dogs were assessed prior to training at Lackland Air Force Base (hereafter termed 'test 3'). Sample sizes of dogs for analyses were variable because dogs that delivered very low scores on an individual behavioral item or a cluster of items during tests 1 or 2 were considered by DoD personnel to have failed the evaluation. Testing of these animals was immediately discontinued, they were not purchased, and they did not contribute any more data after the behavioral item

that resulted in failure. Similarly, some dogs that passed all assessment at test 1 were deemed by DoD veterinarians medically unfit for purchase, and they did not further contribute to the data set. In contrast, test 3 was administered in full to all available subjects, regardless of low scores during testing, and all subjects subsequently entered training.

At all three test times subject dogs were handled and tested by trained DoD personnel with whom the dogs were not familiar. Dogs' reactions were recorded in tests by at least one, and often two, expert observers who served as judges. Behavioral items were scored by judges based on a rating scale which ranged from one to six for each item (Supplementary data, Appendix A). Judges were four working dog practitioners all with greater than 5 years experience in the use of conventional dog behavioral tests. Judges were trained in the application of the ratings system, and were instructed not to discuss their ratings with one another. In general, higher scores on behavioral test items were considered to indicate more appropriate behaviors.

Within each test period (1, 2, or 3) individual items were assessed between 07:00 and 15:00 h over 1 to 2 days, but all items of environmental sureness and detection domains were always assessed indoors and on the same day, while items related to the domains gun sureness and patrol were always assessed outdoors, and in a contiguous manner. The intervals between tests 1, 2, and 3 varied for individual dogs, but ranged from 1 to 7 days between tests 1 and 2, and from 30 to 150 days between tests 2 and 3.

Indoor testing began by walking the dog into the test building, which varied from one location to the next. Most buildings featured a set of stairs, with five to 20 steps leading to a second floor. The buildings normally had slick, concrete floors, and tended to be cluttered with furniture, machinery, and other objects that provided locations for hiding detection "training aids" that provided a target odor. From 5 to 10 people with whom the dogs were not familiar were normally present, recording data (judges), observing (dog vendors and other DoD personnel), or working with the dogs (testers). At test 3, the dogs were completely unfamiliar with the test environment, but at earlier test occasions we were unable to determine how much experience the dogs had with the test facilities. Outdoor testing began by walking the dog into an open outdoor area to a location approximately 10 m from a blind (most often a tree or shrub) where a human tester was hidden, equipped with a plastic and jute fabric bite sleeve and a nylon-handled agitation whip with a 1 m leather lash.

2.3. Specific individual behavioral item test procedure

2.3.1. Environmental sureness

Subject dogs were walked about by the DoD handler on leash for 1 to 5 min, and observed while they investigated and reacted to the environment. At one point in the item test, a loud noise was produced by dropping a heavy object (e.g., a wooden pallet) onto the floor within 2 m of the dog. Most of the dogs were also required to walk up and down five to 20 stairs. Dogs that performed active investigation of the environment/hidden objects, presented objects, and displayed confidence (i.e., carrying the head, ears, and tail high) were scored high, while dogs that were indifferent, apathetic, or performed 'slinking' behaviors with a low head, ears, and tail while shying away or avoiding stimuli were scored low.

2.3.2. Static object interest

Subjects were held by the handler on a 1–1.5 m leash in an open area. A tester approached the dog within 2 m, and drew the dog's attention to a rubber "kong™" toy attached to a rope by whirling the kong in the air for 3–10 s. Testers then placed the kong on the floor within 2 m of the dog, and withdrew. In most cases, the subject dog responded to the sight of the kong with strong-leash pulling and barking. After 5–10 s, testers dropped an object attached to a rope on the ground, rattled it, and then drew the object back. The dropped object was usually a galvanized metal bucket with chains attached, but other objects that made loud noises were sometimes used. If the dog exhibited fear or avoidance of the dropped object, testing was discontinued. If the dog continued to exhibit interest in the kong, then for 5–15 s the intensity of the distraction was increased by repeatedly dropping/throwing the object and dragging it back. Behaviors such as vigorous vocalizations, intense leash-pulling and intense interest in the kong resulted in higher scores in this test, while lack of focus on the kong, trotting around the handler, or breaking off interest when the distracting object was employed resulted in lower scores.

2.3.3. Thrown object interest

Immediately after the static object interest item test, testers recovered the kong and threw it 7–10 m into an open space. The handler waited until the kong came to rest, and then released the subject dog. Dogs were scored according to the rapidity and vigor with which they approached and made contact with the kong. Behaviors such as running full speed after the kong, pouncing and biting it vigorously resulted in high scores, while trotting non-chalantly after the kong, inspecting it or bumping it carefully prior to picking it up resulted in lower scores.

2.3.4. Possession

Immediately after the thrown object interest item test, handlers recalled the subject dog and allowed it to interact for 30–60 s with the kong without instructions. Behaviors such as lying down and chewing vigorously on the kong, or repeatedly dropping or throwing it and then scrambling to catch it again resulted in high scores. Behaviors such as dropping the kong and not hurrying to recover it, or becoming distracted away from the kong by objects or people in the test area, resulted in lower scores.

2.3.5. Physical possession

Immediately after the possession item test, testers took hold of the short rope attached to the kong and attempted to take it away from the dog. If at any point testers were successful in wresting the kong away, he/she allowed the dog to seize the kong again and then immediately re-applied traction. This procedure was then repeated two to four times, at which time the dog was allowed to take the kong and play with it for a few moments. The handler then

gained possession of the kong by lifting upwards on the dog's collar, using a hand to pry the kong out of the dog's mouth, or giving the "out!" release command. Dogs that did not release the kong, and dogs from whom even the handler had great difficulty recovering the kong, received higher scores. Dogs that allowed the kong to be repeatedly pulled from their mouths with modest force received lower scores.

2.3.6. Search activity

Search activity item tests began with testers enticing subject dogs with the kong and then attempting to 'trick' the dog by pantomiming placing the kong in a hiding place 3–4 m from where the dog was held on leash by handlers. The hiding place contained a training aid of one of three substances (sodium or potassium chlorate, cannabis, or smokeless gunpowder). Hiding places were most commonly in drawers/cupboards and were no more than 70 cm from the ground. After pantomiming placing the kong, testers stepped away from the hiding place and the handler released the dog to search. When subject dogs found and exhibited an orienting response to the training aid, they were rewarded with the kong, and allowed to interact with it for a brief period while the tester moved the training aid to another hiding place 4–7 m away from the initial one. The handler then recovered the kong from the dog, the tester again pantomimed placing it in the initial hiding place, and the subject dog was released. A third and sometimes fourth and fifth search were conducted in exactly the same manner. Dogs that received high scores in this test searched strongly without handler input, displayed vigorous, uninterrupted search over long durations, and used olfaction rather than vision in their search behavior. Dogs that received lower scores in this test searched lackadaisically and with interruptions, required handler input/encouragement to maintain search behavior, and/or appeared to rely on vision instead of olfaction. Scoring was based on the stubbornness, concentration, and vigor with which the dogs searched, rather than on how adept they were at recognizing training aid odor and localizing it.

2.3.7. Search stamina

Search stamina items, like search activity ones, were scored over the course of the series of search exercises described above. Dogs that performed comparatively large amounts and long durations of intensive physical work while searching and interacting with the kong reward, without exhibiting signs of physical fatigue such as heavy panting, were given higher scores for search stamina. Dogs that began panting heavily after comparatively less physical work, and/or that rapidly appeared to become fatigued, received lower scores. Particular attention was paid to the dog's ability to maintain "closed-mouth" sniffing (Moulton and Marshall, 1979; Keiichi and Tucker, 1985). Dogs that engaged in a large work output while maintaining closed-mouth sniffing received the highest scores.

2.3.8. Defense

Testers appeared from behind the blind and approached the dog in a direct, slow, and threatening manner (i.e., stalking towards the dog while staring directly into the dogs' eyes, and pausing or "freezing" for effect). When the tester reached a point approximately 2 m from the dog, he/she paused for a moment, then sprung suddenly at the dog with a threatening vocalization and a menacing gesture of the whip, but stopping short of physical contact. Behaviors such as vigorous barking and lunging, and snapping at the tester's arms and hands, received high scores. Behaviors such as looking away from the tester or retreating from him/her, received lower scores.

2.3.9. Threat aggression

Dogs were also assigned a score for their aggressiveness, based on their degree of postural and expressive threat exhibited towards the tester during defense item tests. Dogs that exhibited guttural vocalizations, continuous barking, displayed their teeth in a snarl (as opposed to simply attempting to bite), and erected the hair on their backs, received higher scores, whereas dogs that showed little interest in the tester or that showed interest and excitement but barked in a shrill/light tone without snarling or piloerection received lower scores.

2.3.10. Non-threat bite quality

Following the defense item test, subject dogs were allowed to make contact with the testers' bite sleeve. Dogs normally leaped at the tester, gripped the bite sleeve, and held on. After bite contact was made, the tester walked several steps around the handler (who retained dog on leash) dragging the dog with him/her, moving calmly with low intensity, and without threatening or harassing the dog. After 10–20 s, the tester dropped the sleeve from his/her arm. Normally, the dog retained its bite. Dogs that contacted the bite sleeve with an audible impact, gripped it with the entire jaw (as opposed to using the canines only) with crushing, continuous force, received high scores. Dogs that hesitated prior to biting, or bit softly or shallowly at first, or bit and let go, or chewed on the sleeve received lower scores.

2.3.11. Threat bite quality

After the non-threat bite quality item test, handlers recovered sleeves, returned them to testers, and testers walked away from the dog to a distance of approximately 7 m. Testers then turned and moved quickly at the dog, cracking the whip and vocalizing loudly, until the dog could re-contact the bite sleeve. Testers then walked one or two paces, dragging the dog, and continued to threaten it. No subject dogs were struck in this test; instead, threats consisted of striking the whip-handle on the leash held by the handler. Whip-threats were repeated two to five times while the dog was on the sleeve, after which testers dropped the sleeve. Dogs that bit strongly, calmly, and quietly received high scores, whereas dogs that hesitated to bite, bit and let go, bit shallowly, or vocalized and released the bite under threat received lower scores.

2.3.12. Attention transfer

Immediately after the threat bite quality item test, testers walked 3 m away from the dog, and then moved towards its side/flank, in an attempt to cause the dog to drop the bite sleeve and initiate pursuit. Initially, testers

attempted to initiate pursuit by moving calmly with a neutral expression, but became more threatening if the dog did not transfer its attention from the sleeve. If/when attention was transferred from the sleeve, testers withdrew slowly and calmly while facing the dog. Transfer scores were given based on how rapidly the dog redirected attention away from the sleeve and towards testers, and how long dogs remained focused on the tester as he/she withdrew. Dogs that released the sleeve immediately after testers dropped it, without requiring any threatening encouragement to do so, lunged towards the tester, and remained focused on the tester without returning to the sleeve as the tester moved away received high scores. Dogs that continued to focus attention on the sleeve received lower scores

2.3.13. Pursuit bite

After an interval of 1 or 2 min for the dogs to rest following the attention transfer item test, testers ran away from the dog and handler, waving/cracking the whip, striking the sleeve with the whip, and making loud verbal noises. At a distance interval of approximately 30 m handlers released dogs to pursue; simultaneously testers continued to move away from dogs but in an unthreatening manner with the bite sleeve held out to the side. Dogs that pursued at top speed and leapt at the tester, made biting contact with audible impact, and bit deeply and with force and without "mouthing" or vocalizing received high scores. Dogs that pursued at less than top speed or deviated from a straight line, that hesitated once within range of the tester, missed the bite, or bit shallowly, softly, or with chewing and vocalization received lower scores.

2.3.14. Frontal bite

Following the pursuit bite item test, handlers recovered the bite sleeve from dogs and returned them to testers, and testers again ran away from the dog and handler, waving/cracking the whip, striking the sleeve with the whip, and making loud noises. After 25 m testers stopped and turned to face the dog with the sleeve held low and to the side or behind his/her back. Handlers released dogs to pursue, and as the dog came close to the tester, he/she rushed at the animal and threatened it verbally and with gestures of hand, bringing the sleeve out in front only at the last moment before the dog made contact. Scores in this item test were evaluated exactly as in the pursuit bite item test.

2.3.15. Gun sureness

Either before or after all patrol domain behavioral item tests were given, testers walked from a distance of about 75 m directly towards the dog and fired rounds of 0.38 caliber blank ammunition into the air with a revolver. Two rounds each were fired at approximately 75 paces, 30 paces and 5 to 10 paces from the dog. Subject dogs were evaluated in terms of their steadiness or sureness. Dogs that remained calm and inquisitive and moved about freely received the highest scores. A certain amount of excitable barking, especially when accompanied by a "neutral" facial expression, was also associated with a moderately high score, so long as the dog remained at the end of the leash and did not retreat from the tester. Shying or startling in response to the gunshots, tail-tucking, moving away from the tester and/or hiding behind the handler, and jumping up against the handler apprehensively all resulted in lower scores.

2.4. Data analysis Aim 1: evaluating inter-rater and test–retest reliability of single item behaviors

To assess inter-rater agreement for each of the 15 single behavioral items (Shrout and Fleiss, 1979), we used a one-way random effects intraclass correlation coefficient (ICC) because the specific observer was not identified in most cases. Many single item behavioral items were only rated by a single observer, so we computed both the average ICC (1, k) and the single-rater ICC (1, 1) for each behavioral item to evaluate whether we could justify using a single observer's rating to generate a larger sample size for subsequent analyses. Due to subject attrition, N was variable for each comparison and is given in Table 1. Single-rater reliability was generally high (Section 3.1) so we used Spearman-rank correlations between time pairs of single-observer item scores to evaluate test–retest reliability. We used non-parametric correlations to evaluate test–retest reliability because variances of items were different at different time points, and several single item behaviors were not normally distributed. Due to subject attrition, N was variable for each comparison and is given in Table 2.

2.5. Data analysis Aim 2: evaluating the use of PCA to generate aggregate trait scores

PCA was based on single-observer ratings at time 1 ($N = 554$) and 3 ($N = 456$). We did not perform PCA on time 2 observer ratings because the recommended minimum cases-to-variables ratio for PCA at this time point was not reached (Tabachnick and Fidell, 1996). Twelve of the 15 single behavioral items were included in both PCAs. The items gun sureness, pursuit bite, and frontal bite were not included because only 14 dogs were rated in pursuit and frontal bite tests at time 1, and because a restricted sample of dogs was measured for gun sureness at both time 1 and 3 (Table 1). Single-rater reliability for gun sureness was also low (0.30) at time 1.

At both time points, orthogonal varimax and oblique direct oblimin rotated solution matrices were examined, and both methods resulted in the same pattern of loadings of single item behaviors; for ease of interpretability and in keeping with widespread practice, we report only orthogonal results here. The number of components extracted for each solution matrix was based on a scree test, evaluations of simple structure, and the interpretability of the components themselves (Cattell, 1966; Zwick and Velicer, 1986). For component interpretation, behaviors with a loading of at least 0.40 were considered to contribute to the meaning of a component (Tabachnick and Fidell, 1996).

We tested the extent to which the factor structure generalized across the two testing occasions using targeted Procrustes rotation (McCrae et al., 1996). Targeted rotation of solution matrices assesses how the pattern of component loadings obtained for single item behaviors at time 1 were replicated at time 3. Congruence coefficients greater than 0.85 were taken as evidence of replication across the two matrices (Lorenzo-Seva and ten Berge, 2006).

Table 1
Inter-rater reliability of single item behavioral scales at three different time points from German shepherd, Belgian malanois, and Dutch shepherd dogs used in the Lackland MWD certification program.

Behavior item	Time 1				Time 2				Time 3			
	ρ (single)	ρ (average)	*N*	*F*	ρ (single)	ρ (average)	*N*	*F*	ρ (single)	ρ (average)	*N*	*F*
Environmental sureness	0.87	0.93	256	14.77*	0.85	0.92	70	12.51*	0.90	0.95	330	19.55*
Static object interest	0.80	0.89	252	8.80*	0.80	0.89	70	9.18*	0.81	0.90	340	9.80*
Thrown object interest	0.66	0.80	250	4.93*	0.64	0.78	70	4.61*	0.76	0.87	338	7.47*
Possession	0.74	0.85	250	6.77*	0.77	0.87	69	7.90*	0.72	0.84	339	6.17*
Physical possession	0.78	0.88	248	8.25*	0.78	0.88	70	8.21*	0.89	0.94	338	16.79*
Search activity	0.79	0.88	251	8.41*	0.78	0.88	70	8.23*	0.79	0.89	336	8.73*
Search stamina	0.55	0.71	244	3.50*	0.52	0.69	68	3.19*	0.64	0.78	332	4.56*
Defense	0.86	0.93	224	13.68*	0.65	0.79	70	4.75*	0.87	0.93	306	14.77*
Threat aggression	0.64	0.78	208	4.55*	0.57	0.72	62	3.64*	0.69	0.82	305	5.46*
Non-threat bite quality	0.73	0.84	206	6.46*	0.75	0.86	69	7.11*	0.88	0.94	306	15.65*
Threat bite quality	0.73	0.85	204	6.50*	0.78	0.88	69	8.15*	0.87	0.93	304	14.02*
Attention transfer	0.68	0.81	202	5.35*	0.56	0.72	68	3.57*	0.76	0.86	305	7.33*
Pursuit bite	0.27	0.42	14	1.73	0.88	0.94	69	15.59*	0.87	0.93	305	14.66*
Frontal bite	0.40	0.57	14	2.35	0.82	0.90	68	10.02*	0.87	0.93	302	14.10*
Gun sureness	0.30	0.47	140	1.88*	0.87	0.93	10	13.89*	0.84	0.91	194	11.49*

* $P < 0.0001$.

Quantitative analyses of the similarity of single item loadings across the two PCA solutions were high (Section 3.2), so in order to quantify aggregate behaviors we generated scores based on the pattern of loadings that were obtained. All behavioral items were measured on the same scale (one to six), so we calculated average unit-weighted scores for each time period (time 1, 2, and 3) by averaging the single-observer ratings given to a dog for a particular component identified in PCA. For example, two behaviors that clustered with themselves but not with other variables at times 1 and 3 were 'search activity' and 'search stamina'. Thus, in order to create an aggregate score for the broader dimension, which we called 'search focus', we averaged the single-observer rating given in the search activity and search stamina tests at a single time point. We computed separate component scores for each dog for each component at each time point, resulting in 12 unique scores per dog (four component scores per time point, three time points). Average unit weighting was used instead of regression methods to generate scores to facilitate future attempts at independent study validation (Gorsuch, 1983; Goldberg and Digman, 1994).

We used Spearman-rank correlations between time pairs of aggregate scores to evaluate test–retest reliability through time. To evaluate whether the same test–retest effects were obtained using aggregate scores and using single item ratings, we assessed the patterns of means of correlation coefficients between aggregate scores and between single behavioral item ratings using Fisher's *z* scores and weighting means by sample size; All *N*'s are given in Table 2.

2.6. Data analysis Aim 3: prediction of certification outcomes

We used a series of binary logistic regression models to predict certification outcomes using an individuals' behavior, breed, and sex. In all models, certification outcomes (pass, fail) were our dependent variable of interest. The central goal was to evaluate whether we could predict certification using behavior at the time of procurement (time 1) or at the start of training (time 3), so test time-specific models were used (i.e., two sets of models using time 1 and time 3 data separately). Our sample consisted of a total of 357 dogs that received patrol testing certification tests and 356 dogs that received odor-detection certification tests. Of these, 63.9% of dogs were patrol certified, 76.5% of dogs were detection certified, and 60.2% of dogs were dual-certified.

Given the exploratory nature of our analyses, we concentrated on a model's prediction error (PE) instead of focusing on parameter *P* values to evaluate model usefulness. We first randomly chose 70% of the certification subjects to serve as our training data set (i.e., data from which model parameters were estimated) and 30% of certification subjects as our test data set (i.e., data from which fitted parameters were tested to predict to new cases). PE for each model was assessed by the percentage of cases in the test data that were correctly classified using the model parameters generated by the training data. Previous to fitting models, we examined Spearman-rank coefficients

Table 2
Spearman-rank correlation coefficients through time for component scores and for single item behaviors.

	Time 1 to Time 2	Time 2 to Time 3	Time 1 to Time 3	Weighted mean
Single item behaviors				
Environmental sureness	0.52 (72,<0.001)	0.54 (61, <0.001)	0.16 (458, <0.001)	0.33
Static object interest	0.59 (76, <0.001)	0.50 (60, <0.001)	0.35 (498, <0.001)	0.48
Thrown object interest	0.47 (76,<0.001)	0.18 (60, 0.18)	0.20 (497,<0.001)	0.27
Possession	0.57 (76,<0.001)	0.45 (60,<0.001)	0.24 (498,<0.001)	0.38
Physical possession	0.51 (76,<0.001)	0.42 (60, 0.001)	0.30 (497,<0.001)	0.41
Search activity	0.43 (75,<0.001)	0.17 (58, 0.20)	0.20 (495,<0.001)	0.26
Search stamina	0.47 (72,<0.001)	0.17 (53, 0.23)	0.22 (492,<0.001)	0.28
Defense	0.63 (71,<0.001)	0.36 (60, 0.005)	0.26 (491,<0.001)	0.38
Threat aggression	0.77 (61,<0.001)	0.56 (52,<0.001)	0.27 (488,<0.001)	0.44
Non-threat bite quality	0.59 (70,<0.001)	0.43 (59, 0.001)	0.28 (490,<0.001)	0.40
Threat bite quality	0.56 (70,<0.001)	0.44 (60,<0.001)	0.25 (490,<0.001)	0.38
Attention transfer	0.66 (69,<0.001)	0.60 (56,<0.001)	0.31 (485,<0.001)	0.47
Pursuit bite	Data unavailable	0.59 (60,<0.001)	0.07 (18, 0.77)	0.53
Frontal bite	Data unavailable	0.55 (59,<0.001)	0.19 (18, 0.44)	0.58
Gun sureness	0.50 (14, 0.07)	0.64 (22, 0.001)	0.08 (335, 0.14)	0.17
Aggregate component scores				
Object focus	0.58 (72,<0.001)	0.43 (60, 0.001)	0.30 (454, 0.001)	0.43
Sharpness	0.63 (69,<0.001)	0.51 (59,<0.001)	0.29 (489,<0.001)	0.43
Human focus	0.68 (60,<0.001)	0.49 (49, <0.001)	0.29 (480, <0.001)	0.42
Search focus	0.49 (72, <0.001)	−0.14 (53, 0.32)	0.20 (491, <0.001)	0.21

Sample size and *P* values are given in parentheses. Object focus score = Environmental sureness + Static object interest + thrown object interest + possession + physical possession. Sharpness score = Defense + Non-threat bite quality + Threat bite quality. Human focus score = Defense + Threat aggression + Attention transfer. Search focus score = Search activity + Search stamina.

between behavioral predictors to assess inter-correlations. At time 1, no correlations between the four aggregate component scores were greater than 0.44, and at time 3, no correlations between the four aggregate scores were greater than 0.48. For single item behaviors, at both time 1 and time 3, threat bite and non-threat bite quality scores tended to be positively correlated (time 1: Spearman's $r_{(348)} = 0.63$; time 2: Spearman's $r_{(336)} = 0.79$); outside of this pair-wise combination the highest absolute pair-wise correlation was 0.52. Therefore, for fitted models using single item behaviors, we dropped values for non-threat bite. We ran six separate main-effects logistic regressions at each time point:

1) Model 1: dual-certification outcome = aggregate score 1 + aggregate score 2 + aggregate score 3 + aggregate score 4 + sex + breed (two levels: DS were classified as BM as there were only 22 of the former)
2) Model 2: patrol certification outcome only = aggregate score 1 + aggregate score 2 + aggregate score 3 + aggregate score 4 + sex + breed
3) Model 3: detection certification outcome only = aggregate score 1 + aggregate score 2 + aggregate score 3 + aggregate score 4 + sex + breed
4) Model 4: dual-certification outcome = environmental sureness score + static object interest score + thrown object interest score + possession score + physical possession score + search activity score + search stamina score + defense score + threat aggression score + threat bite quality + attention transfer + gun sureness + sex + breed
5) Model 5: patrol certification outcome only = the 11 single item behaviors indicated in Model 4 + gun sureness + sex + breed
6) Model 6: detection certification outcome only = the 11 single item behaviors indicated in Model 4 + gun sureness + sex + breed.

We also included the single items 'pursuit bite' and 'frontal bite' in time 3 models since inter-rater reliability was high and sample sizes were sufficient. Once we identified, based on PE, the best model above, we then fitted a reduced model based on important predictors to determine which single item tests, if any, might be able to be dropped in future behavioral pre-screening protocols. For each individual model, we assessed overall model significance based on the log-likelihood ratio test (i.e., compared the fully parameterized model with the model containing only a fitted constant), and after fitting models we examined residuals and multicollinearity diagnostics to satisfy the assumptions of logistic regression (Field, 2005). There were no outliers or influential cases identified, and multicollinearity assumptions were also satisfied for each model. For time 1 Models 1-2 training data *N* was 235, and test *N* was 97. For time 1 model 3, test *N* was 98. For time 1 Models 4-5 training *N* was 215, and test *N* was 88. For time 1 model 6, test *N* was 89. For time 3 Models 1-2 training *N* was 222, and test *N* was 89. For time 3 Model 3, test *N* was 88. For time 3 Models 4-6 training *N* was 164, and test *N* was 66. For reduced models, *N* is given in Section 3.3. We used SPSS 15.0 for all analyses.

2.7. Ethics

Procurement of dogs and experimental methods were approved by the University of Texas Institutional Animal Care and Use Committee (IACUC protocol 07092902). Once procured from vendors, dogs were under the care of the United States Army's Veterinary Command. On a monthly

basis, the dogs were weighed, their physical condition and diet were evaluated, and they were treated with drugs to control parasites. In addition, dogs received diagnostic radiology and blood testing, dental cleaning, vaccinations against rabies and other viral diseases, and all necessary clinical and surgical veterinary treatment to ensure their health and physical welfare. Female dogs were spayed after their arrival at Lackland but before test 3. Male dogs were normally intact. After procurement, all dogs that were eliminated were donated to civil law-enforcement agencies or to private homes.

3. Results

3.1. Inter-rater and test–retest reliability of single item behaviors

At all three test periods, inter-observer correlation coefficients for single item ratings were high. Unit-weighted average reliability (using Fisher's r to z transformations) at time 1 was 0.69 for a single observer and 0.81 for the average observer ratings, 0.75 for a single observer and 0.86 for average observer ratings at time 2, and 0.82 for a single observer and 0.90 for average observer ratings at time 3. Within this high overall reliability there were discrepancies, but only during testing at time 1. Two single items at time 1, gun sureness and pursuit bite, had lower average and single-rater reliability (<0.47), while one other item at time 1 (frontal bite) had low single-rater reliability (0.40). All other single items had single and average reliabilities >0.52 (Table 1).

Table 2 (top panel) gives Spearman-rank correlation coefficients of the pair-wise comparisons through time for the 15 single item behaviors. In general, test–retest reliability was strong across the shortest time period (time 1 to 2), with all coefficients nearing or exceeding 0.50 except for search activity item ratings (Spearman's $r = 0.43$). Over longer time periods (times 2 to 3 and 1 to 3), there was a decrease in absolute value of rank coefficients, with the lowest coefficients observed over the longest time span (time 1 to time 3). Exceptions to this general pattern, however, occurred for thrown object, search activity, and search stamina item tests, where coefficients were small and similar across times 2 to 3 and across tests 1 to 3.

3.2. PCA solution matrix validity and test–retest reliability of component scores

Four components from the PCAs at time 1 and 3 were chosen as a best fit of the data (56.6% of the variance accounted for at time 1; 65.0% of the variance accounted for at time 2; Table 3). The component names were chosen based on the definitions of the items that loaded strongly on each component and, where possible, trying to use terms consistent with previous research (reviewed in Jones and Gosling, 2005; Graham and Gosling, 2009). At both time periods, the same single item ratings loaded strongly on the same components. The first component, which we named object focus, described dogs that differed in their steadiness in environmental sureness tests and in their attention to objects in static and thrown object tests and possession and physical possession tests. The second component described dogs that varied with respect to their willingness to respond with an aggressive response, or their sharpness (Wilsson and Sundgren, 1997b), with some pulling vigorously on the handler's leash when confronted with a threat, and making intense, focused biting contact in non-threat and threat bite tests. Others exhibited disinterest towards threat stimulation in defense aggression tests, and did not bite or bit only quickly and distractedly in non-threat and threat bite quality tests. Similarly, component 3 described a continuum of dogs that varied with respect to their levels of human focus, with some dogs exhibiting intense aggressive threats during testers' approach (e.g., tooth exposure and piloerection) and exerting intense effort to approach and contact the tester, and others that attempted to ignore or recoiled from the presence and approach of the tester. Finally, component 4 described variation between dogs in their overall search focus. Dogs on the upper end of this continuum exhibited vigorous sniffing and searching behavior with a high level of attention in search tests, and maintained this high level of activity in search stamina ones. Dogs on the lower end of the search focus continuum exhibited interrupted or feeble search behavior while searching, and tended to tire/lose interest quickly in repeated search tests.

In the targeted Procrustes rotation of time 1 item loadings to those obtained at time 3 the overall factor congruence was excellent (0.97) and the individual component congruences were similarly high for object focus (0.97), sharpness (0.97), human focus (0.98), and search focus (0.97), indicating a high degree of convergence for all four components in our data set. In addition to single item behaviors loading on the same components across times 1 and 3, single item behaviors that loaded on a particular component at time 1 tended not to load on a different component at time 3 (Table 4).

Table 2 (lower panel) gives Spearman-rank correlation coefficients of the pair-wise comparisons through time for the four aggregate scores. Similar to the pattern observed for single item scores, there was a decrease in absolute value of component score rank coefficients through time, with the lowest coefficients observed over the longest time span (time 1 to time 3). Spearman-rank coefficients were moderate to high across times 1 and 2 and across times 2 to 3 (0.4 to 0.7), with the exception of 'search focus' across times 2 to 3 (Spearman's $r = -0.14$). However, this lack of correlation was also observed with the two single item behaviors (search activity and search stamina) that went into making the 'search focus' score. Across time 1 to time 3, coefficients for component scores were small (<0.30), indicating that rank-order behavior in our sample was generally maintained across shorter time periods, but not necessarily so across longer ones. For search focus scores, there was no evidence of test–retest reliability from time 2 to time 3 and from time 1 to 3. In general, qualitative comparison between single item and aggregate score correlations indicated that test–retest reliability coefficients for single item estimates were similar to the patterns observed using aggregate scores.

Average correlation coefficients using aggregate scores were higher than average coefficients for single item

Table 3
Component loadings of behavioral ratings at times 1 and 3 on four orthogonally rotated principal components. Boldface indicates the highest component loading(s) for each behavior.

Principal components time 1				
Behavioral item	Object focus	Sharpness	Human focus	Search focus
Environmental sureness	**0.62**	−0.06	0.12	−0.17
Static object interest	**0.73**	0.02	0.11	0.09
Thrown object interest	**0.68**	0.12	−0.08	0.23
Possession	**0.69**	0.05	−0.06	0.17
Physical possession	**0.39**	0.23	−0.11	0.25
Search activity	0.23	0.09	0.03	**0.75**
Search stamina	0.03	−0.06	0.12	**0.79**
Defense	−0.01	**0.41**	**0.55**	0.01
Threat aggression	0.20	0.07	**0.74**	0.08
Non-threat bite quality	0.16	**0.88**	0.08	−0.06
Threat bite quality	−0.01	**0.87**	0.00	0.10
Attention transfer	−0.13	−0.12	**0.72**	0.05
% variance explained	22.2	13.9	11.4	9.1
Principal components time 3				
Environmental sureness	**0.60**	−0.16	0.25	−0.25
Static object interest	**0.75**	0.17	0.01	0.12
Thrown object interest	**0.76**	0.13	0.06	0.10
Possession	**0.76**	0.16	−0.01	0.19
Physical possession	**0.62**	0.24	−0.08	0.12
Search activity	0.32	0.17	−0.02	**0.72**
Search stamina	0.03	−0.12	0.07	**0.82**
Defense	0.23	**0.43**	**0.67**	−0.10
Threat aggression	0.06	0.23	**0.75**	−0.07
Non-threat bite quality	0.20	**0.90**	0.11	−0.01
Threat bite quality	0.18	**0.89**	0.12	0.04
Attention transfer	−0.13	−0.20	**0.68**	0.23
% variance explained	29.9	14.4	10.9	9.7

Note: Order of behavioral items above is listed according to the order presented in Section 2.3, and approximated the actual order of tests given.

behaviors across times 1 and 2 ($t_{(588.14)} = 5.42$, $P < 0.001$). However, the absolute mean difference between the two average coefficients was small (aggregate scores' average Spearman's $r = 0.59$; single item behaviors' average Spearman's $r = 0.57$). Average correlation coefficients using single item behaviors were higher on average across times 2 and 3 ($t_{(267.13)} = -5.73$, $P < 0.001$), but the absolute difference was again small (aggregate scores' average Spearman's $r = 0.34$; single item behaviors' average Spearman's $r = 0.44$). Across times 1 to 3, aggregate score average correlations were higher than single item correlation averages ($t_{(4936.06)} = 20.35$, $P < 0.001$), but the aggregate score coefficient average Spearman's r was 0.27 and the single item coefficient average Spearman's r was 0.24. We interpreted these results as indicating that test–retest reliability was not affected by our use of aggregate scores.

3.3. Prediction of certification outcomes

At time 1, none of the six models predicted any final certification outcome better than the constant-only model where all dogs were assumed to pass (Model 1: $\chi^2_{(6)} = 5.82$, $P = 0.44$; Model 2: $\chi^2_{(6)} = 10.73$, $P = 0.10$; Model 3: $\chi^2_{(6)} = 7.21$, $P = 0.30$; Model 4: $\chi^2_{(14)} = 11.92$, $P = 0.61$; Model 5: $\chi^2_{(14)} = 12.95$, $P = 0.53$; Model 6: $\chi^2_{(14)} = 8.06$, $P = 0.89$). PE to test data for the full models versus constant-

Table 4
Targeted Procrustes rotation comparing single item behavioral component loadings obtained at time 1 to those obtained at time 3. Boldfaced loadings indicate the item the factor loaded on in the time 1 sample.

Principal component					
	Object focus	Sharpness	Human focus	Search focus	Item congruence
Environmental sureness	**0.59**	−0.10	0.15	−0.22	0.99
Static object interest	**0.73**	−0.01	0.14	0.03	0.95
Thrown object interest	**0.71**	0.10	−0.05	0.15	0.99
Possession	**0.70**	0.03	−0.03	0.10	0.98
Physical possession	**0.43**	0.23	−0.10	0.20	0.97
Search activity	0.30	0.12	0.02	**0.72**	1.00
Search stamina	0.10	−0.02	0.09	**0.79**	0.99
Defense	−0.02	**0.40**	**0.56**	0.01	0.95
Threat aggression	0.17	0.04	**0.75**	0.09	0.94
Non-threat bite quality	0.19	**0.86**	0.12	−0.12	0.99
Threat bite quality	0.04	**0.88**	0.03	0.05	0.98
Attention transfer	−0.16	−0.13	**0.70**	0.09	0.98
Factor congruence	0.97	0.97	0.98	0.97	0.97

only models was the same in Models 1-3 (all models using aggregate scores) and 5-6 (models using single item scores). For Model 4, there was a slight improvement in PE using the full model (58% of independent cases correctly classified) compared to the constant-only model (56.8% of cases), but this improvement, as indicated above, was not better than chance.

At time 3, models 1, 2, and 3 using aggregate scores all fit training data better than constant-only models (Model 1: $\chi^2_{(6)} = 20.03$, $P = 0.003$; Model 2: $\chi^2_{(6)} = 22.79$, $P = 0.001$; Model 3: $\chi^2_{(6)} = 17.19$, $P = 0.009$). However, only in models 1 and 2 was PE to test data improved using the full model (for model 3, PE to test data decreased from 75.6% for the constant-only model to 73.3% for the model with behavior, breed, and sex). For model 1, PE to test data was 66.3%, compared to 59.6% for the constant-only version, and only one aggregate behavior, search focus, was significant in predicting dual-certification outcome. Odds of successful dual-certification were increased 2.59 times with a one unit increase in search focus scores (Wald's $\chi^2_{(1)} = 11.09$, $P = 0.001$). For model 2, PE to test data was 66.3%, a 1.4% increase over PE from the constant-only version. Search focus (Wald's $\chi^2_{(1)} = 9.84$, $P = 0.002$) and sharpness (Wald's $\chi^2_{(1)} = 4.58$, $P = 0.03$) scores were identified as the important predictors for patrol certification, with a 2.47 increase and a 1.71 increase in the odds of successful certification, respectively, with a one unit increase in aggregate behavior. A reduced model predicting dual-certification outcome using only search focus and sharpness scores resulted in a PE of 64.2%, while the same reduced model predicting patrol outcome resulted in a PE of 66.3%, indicating that dropping all other aggregate behaviors (and therefore the single item measurements that went into generating them) would result in a slight worsening in PE with regards to predicting dual-certification predictions, but would not influence patrol certification prediction at all.

At time 3, models 4 and 5 using single item behaviors also fit training data better than constant-only models (Model 4: $\chi^2_{(14)} = 45.23$, $P < 0.001$; Model 5: $\chi^2_{(14)} = 42.99$, $P < 0.001$), but not for model 6 ($\chi^2_{(14)} = 21.22$, $P = 0.17$). However, in the case of model 4 (dual-certification), PE to test data was slightly worse for the parameterized model (60.6% PE compared to 62.1% PE for constant-only version). For model 5 (patrol certification) PE to test data was 71.2%, a 4.5% improvement over the constant-only version. Frontal bite was the only single item at time 3 to reach significance, (Wald's $\chi^2_{(1)} = 7.04$, $P = 0.008$), with a 2.76 increase in the odds of successful patrol certification with a one unit increase in the frontal bite scale. The significance of static object interest and search stamina items were borderline (static object interest: Wald's $\chi^2_{(1)} = 3.59$, $P = 0.06$; search stamina: Wald's $\chi^2_{(1)} = 3.33$, $P = 0.07$), so we fit a reduced patrol certification model at time 3 with frontal bite, search stamina, and static object interest only (training $N = 232$, test $N = 94$). PE in this reduced model was 73.4%, a 7.4% improvement over the constant-only version. Once again, frontal bite was the only single item in the reduced time 3 model to reach significance, (Wald's $\chi^2_{(1)} = 18.18$, $P < 0.001$), with a 2.20 increase in the odds of successful patrol certification with a one unit increase in the frontal bite scale. A further reduced model with frontal bite only (training $N = 239$, test $N = 95$) had a PE of 68.4, a 2.1% improvement over the corresponding constant-only version.

In summary, information regarding dog breed and sex did not improve prediction of any certification outcome at either time 1 or time 3, and information regarding dog behavior at time 1 also did not improve prediction errors. At time 3, information on behavior did not improve prediction of odor-detection certification, but aggregate behavior scores, mainly search focus and sharpness, improved PE by 6.7% for dual- and 1.4% for patrol-only certification outcomes. Frontal bite, search stamina, and static object interest were the only single item behaviors to improve prediction error of certification outcomes at time 3, but only for patrol certification, improving PE from 2 to 7%.

4. Discussion

4.1. Summary

Understanding personality variation between dogs and how these dispositions relate to important working domain tasks remains an outstanding issue for working dog programs whose mission it is to procure and train dogs (Goddard and Beilharz, 1983; Slabbert and Odendaal, 1999; Svartberg, 2002; Rooney et al., 2007). Here we present the results of a first exploratory analysis of the reliability and validation of a behavioral test instrument designed to predict MWD certification in military patrol and odor-detection working environments. Overall, we found that the measurement instrument currently in use by the USA Department of Defense shows very high inter-rater reliability for ratings of single item behaviors. Quantifying aggregate behaviors was also a useful tool to summarize MWD behavior. PCA analysis indicated that 'object focus', 'sharpness', 'human focus', and 'search focus' summarized large portions of behavioral variation among dogs. Importantly, the generalizability of the meaning of these four personality traits was also strong, indicating that the pattern of inter-relationships between single item behaviors was stable. Aggregate scores showed comparable levels of test–retest reliability compared to single item behaviors, and predictive models that used aggregate scores rather than single item behaviors tended to be more powerful predictors of patrol and dual-certification outcomes, but mainly using search focus and sharpness scores measured at time 3. Frontal bite, search stamina, and static object interest tests were the only single item behaviors that successfully improved prediction of certification outcomes, but only for patrol certifications. Below, we discuss the implications of these results for patrol and detection MWD programs.

4.2. Reliability of behavioral measurement instruments

Our results indicate that the first reliability criterion is satisfied for the Lackland measurement instrument, since average inter-rater reliability at all three time periods for 12 out of the 15 single item behaviors was extremely good (overall mean coefficient = 0.86). Gun sureness, pursuit and frontal bite on the other hand, did not indicate suffi-

cient inter-rater reliability at time 1. Pursuit and frontal bite reliability estimates, however, may have been hampered by small sample sizes at time 1 ($N = 14$), as these two single items showed sufficiently high single and average reliability at times 2 and 3 when sample sizes were larger ($Ns > 68$, coefficients > 0.82). These results fit well with what we know from other studies, where inter-rater reliability can be high for most items but not for others. Rooney et al. (2007) found a high level of agreement among 6 independent dog trainers and scientists on 10 aspects of performance-related behaviors relevant to searching in Labrador retrievers (coefficients > 0.52), while one behavior ('responsiveness') had slightly lower reliability (0.49). In a large sample of German Shepherd Dogs used for breeding purposes, Ruefenacht et al. (2002) found that differences between multiple observers in their ratings of eight different behavioral scores accounted for a large amount of the variance in breeding models; 45–125% of the standard deviation of behavior scores were observed between the two most disparate judge ratings for each trait. This was despite the fact that observers were trained for 3 months previous to giving ratings (Ruefenacht et al., 2002). Similar to human personality research (John and Robins, 1993), some behavioral items in animals appear to be more reliably rated by observers than others (Gosling, 2001). Differences between observer ratings can be influenced by various factors including different levels of observer acquaintance with subject animals/test situations and differences in the observability of different traits (for additional explanations, see Gosling, 2001). While we are unable to identify the specific reason for the low inter-rater reliability for gun sureness at time 1 here, our results reinforce the idea that reliability must be tested, and not assumed, because using single item behaviors with low reliability in subsequent analyses could lead to spurious results. Current reports on the criterion of inter-rater reliability of working dog behavior are scarce, and in some cases single-observer ratings without evaluating observer reliability are used (e.g., Wilsson and Sundgren, 1997a,b). Future studies evaluating the properties of measurement instruments, especially as they relate to sensory modalities of dogs and subsequent human perception could provide valuable insight on this matter.

The second core aspect of reliability, test–retest reliability, was assessed here at two levels, one using single item behaviors and one using aggregate trait scores derived from PCA. One of our aims of assessing these two levels was to evaluate whether using aggregate scores resulted in different patterns of test–retest reliability relative to using single item behaviors. Across shorter time periods (time 1 to time 2), we found that on average, most single item behaviors and personality traits showed moderate to high levels of test–retest reliability (coefficients > 0.5). The two single items related to search focus, 'search activity' and 'search stamina', as well as search focus itself had lower levels of reliability across times 1 to 2 (coefficients 0.4–0.5). With increasing amounts of time between test occasions (tests 2 to 3 and tests 1 to 3), however, reliability coefficients tended to decrease, and for search focus and its two single item behavioral indicators, were not different from zero.

Test–retest reliability can be used synonymously with behavioral consistency across time, as both terms can refer to how much individuals, on average, maintain or change their behavior relative to others through time. Reports on the consistency of behavioral traits in patrol or odor-detection dogs are scarce. However, our test–retest reliability results generally support studies from companion and guide dogs where behavioral consistency declines with greater time intervals between samples, and consistency for some behaviors over the same time interval may be different than for others. For example, Svartberg et al. (2005) reported a high degree of consistency (range: 0.57–0.89) for the six personality traits taken from the Swedish dog mentality tests (Svartberg and Forkman, 2002) across a time period of two months, but in another sample across a longer time period (1 to 2 years), consistency was much lower (range: 0.12–0.36) for the same six traits (Svartberg, 2005). In another instance over a 6 month period, Netto and Planta (1997) report a high degree of consistency of aggression towards people from individual companion dogs from several breeds (range: 0.52–0.77), but over the same time period, Goddard and Beilharz (1985) found that aggression towards conspecific dogs was not consistent in individual guide dogs (mean repeatability: 0.20).

Most, if not all, behavioral traits are at least somewhat sensitive to environmental changes, and understanding environmental conditions that influence behavioral change is essential for designing housing and training conditions that maximize and reinforce appropriate working dog behavior (Rooney et al., 2009; Wells, 2009). One potential explanation for the small to moderate effect sizes for consistency between times 2 and 3 was that the transportation of dogs to Lackland, and their subsequent housing on-site, influenced behavior in unknown ways. Stressful situations, in particular, can induce hormonal and behavioral cascades that can impact subsequent long-term behavioral processes (Beerda et al., 1997). Along with this idea of a generalized stress effect on behavior, there is also evidence that some personality types in animals are more likely to change their behaviors than are others (Bell et al., 2009). For example, in laboratory rodents, aggressive individuals tend to display rigid, unchanging behavior in the face of changing environmental conditions, but non-aggressive individuals are more plastic and less routine-driven when laboratory conditions are changed (Benus et al., 1987; Koolhaas et al., 1999). Currently, nothing is known concerning the behavioral and physiological impacts of the current methods used to procure, transport, and house dogs in the Lackland program. Future research on the consistency of behavior and the factors that may influence behavioral change is highly relevant to the Lackland as well as other working dog programs that use behavior as a metric for the selection of individuals for working service.

One potential alternative explanation is that our estimates of test–retest reliability were reflective of differences in observer inter-rater reliability through time, rather than dog behavioral change, since different observers rated the same dogs at all three time points. While the strong single item inter-rater reliability estimates obtained here at each separate time are

reassuring, future attempts by MWD programs using the same observers at all test sites should allow for disentanglement of observer variation versus behavioral change through time. Similar to other working dog programs (e.g., Svartberg, 2005), the MWD program at Lackland is restricted logistically, in that the same observers are often not available to perform all behavioral assays at all test locations. Dog behavior is sensitive to human cues (Kaminski et al., 2009), so an increased focus on training human observers, and adherence to more detailed rules may allow for improvements even in situations where the same observers cannot be used.

Here, we also used our test–retest reliability results to guide our choice of model parameters for prediction of certification outcomes. Specifically, our results suggested that behavior as measured by the Lackland MWD was influenced by test conditions (i.e., tests in Europe versus tests at Lackland), so we fit time-specific models in subsequent tests of predictive validity.

4.3. Predictive validity of behavioral measurement instruments

Improved prediction of training certification outcomes was not obtained by using information on the breed or sex of an individual dog. Breed effects have previously been shown to have a large effect on dog personality, and as such, on suitability for working service (Netto and Planta, 1997; Wilsson and Sundgren, 1997b). However, previous studies that report breed effects used dog breeds that were more distantly related (e.g., Labrador Retrievers versus German Shepherds) than the subjects included in our study (e.g., German Shepherds and Belgian Malanois). Likewise, sex differences have also been reported to influence working outcomes in some dogs (Goddard and Beilharz, 1982; Wilsson and Sundgren, 1997b; Ruefenacht et al., 2002), but not in others (Fuchs et al., 2005). We found no evidence that the sex of an individual significantly influenced its probability of successfully passing certification but this finding could have been driven by the small number of female dogs (22%) in the sample.

In addition, we found that behavior measured previous to procurement in Europe did not predict certification outcomes after training 5–8 months later, supporting the idea that important behavioral changes occurred between testing sites in Europe and prior to training at Lackland. On the other hand, our results indicate that increased search focus and sharpness, as well as higher scores in frontal bite, static object, and search stamina tests after arrival at Lackland successfully improved predictions of dual- and patrol-only certification outcomes. Previous studies also support the idea that differences in police and patrol abilities can be characterized by differences in personality traits such as boldness or aggression (Svartberg, 2002), with successful characterization of certified and uncertified dogs with personality scores reaching 92 and 82%, respectively, in some cases (Slabbert and Odendaal, 1999). While the improvements in prediction observed here were small (2–7%), given the costs of purchasing, importing, housing, and training (approximately $18500US per dog), this small percentage improvement results in a substantial potential savings. It is also worthwhile to note that aggregating single item scores to estimate 'personality' appeared to be a more powerful method than using single item behaviors alone. Heritability estimates of dog personality traits have also been found to be higher than estimates using single item behaviors that went into making the personality scores (Goddard and Beilharz, 1982; Wilsson and Sundgren, 1997a). Use of a personality framework, as long as reliability criterion are satisfied, improves the ease by which the results from several different test situations can be summarized, and may improve prediction of important working outcomes; the latter of course awaits future validation from independent studies.

Somewhat surprisingly, search focus did not predict odor-detection outcomes per se, as our subjective a priori groupings of behavioral items into working domains would have suggested (see Section 2.2 and Supplementary data, Appendix A). Instead, together with a dog's level of sharpness, a dog's search focus behavior predicted dual- and patrol-only certification outcomes. Intuitively, an individual's inherent level of sharpness should influence its ability to successfully perform policing duties. However, the strong relationship of search focus scores and patrol certification outcomes is more puzzling. The identification and measurement of personality traits relevant to odor-detection working domains have rarely been examined (Rooney et al., 2007), but Maejima et al. (2007) found that behavioral items related to amount of activity, obedience, concentration, anxiety, and interest in objects (summarized as the personality trait 'desire to work') successfully characterized successful versus unsuccessful drug detection dogs in Japan. Furthermore, 244 dog handlers and trainers from six UK search dog agencies identified behavioral traits such as tendency to hunt by smell alone, stamina, ability to learn from being rewarded, tendency to be distracted when searching, and motivation to chase an object (amongst others) as the most ideal traits in search specialist dogs (Rooney et al., 2004). One explanation then for our null results regarding predicting odor-detection certification outcomes from search focus scores was that we did not choose the right behavioral items that were closely associated with a dog's ability to pass odor-certification tests. However, it is worth noting that several of the important behavioral items identified above in Maejima et al. (2007) and Rooney et al. (2004) were present in the Lackland MWD measurement instrument (e.g., search activity, search stamina, thrown object interest). Another explanation is that search focus as measured by the Lackland instrument *is* actually related to odor-detection certification outcomes, but the behavioral items used to define this personality trait in the current instrument were not good indicators of a 'search focus' personality trait in our sample of dogs. Indeed, the fact that search focus scores actually predicted dog certification in patrol tests lend some support to this argument. Given the relative paucity of information available regarding the relationship between dog personality traits and odor-detection outcomes, it is clear that further work is needed. For example, a better understanding of the genetic and phenotypic architecture of dog personality traits along with other morphological and physiological traits (e.g.,

Draper, 1995; Svartberg and Forkman, 2002; Ley et al., 2008) as well as increased attention of human cues during odor-detection training and development (Lit, 2009) are likely areas to begin.

In addition to the idea that our behavioral instrument simply did not measure the 'right' personality trait with regards to odor-detection or that it did so in an inefficient manner, another potential explanation for the lack of improved prediction of successful odor-detection certification outcomes (as well as for the small increase in predictive improvement in dual- and patrol certification outcomes) was that current methods used to select dogs for training in the Lackland MWD program have resulted in a relatively high percentage of successfully certified MWDs (60–76% success rates). In other words, we were faced with a restriction of range issue, whereby predictions were hampered by the fact that only dogs that successfully passed behavioral tests at time 1 were given tests at time 2, and only those that passed time 2 tests were given the instrument at time 3. In other words, most dogs that contributed to our predictive models (based on a certification outcome dependent variable) were already high performers. Thus, our estimates of personality traits may actually be stronger predictors of dual- and patrol certification, or odor-detection certification for that matter, but due to the high rates of relative success already achieved in this program (and the corresponding lack of full data for 'failed' dogs), predictive models were unable to improve prediction rates to a greater extent.

4.4. Conclusions

Evaluations of the reliability and validity of behavioral instruments used to select dogs for patrol and odor-detection work are scarce (Graham and Gosling, 2009). This is despite the fact that MWDs are by far and away the most effective and versatile current means of identifying explosives, and anecdotal reports from the field continue to indicate that some dogs are better at their working tasks than are others. Our study provides the necessary evidence for inter-rater reliability for the current measurement instrument in use at Lackland, and also identifies good short-term, but poorer longer-term test–retest reliability. Once at Lackland, improvement in certification rates can be increased by selecting only those dogs with appropriate behavioral profiles with regards to search focus and sharpness, potentially saving significant amounts of staff and monetary resources needed to house and train dogs. Given the increasing need for specialist working dogs in the current worldwide political climate, the significance of evaluating and optimizing methods used to select and train MWDs is clear. Our study is one of the first to provide estimates of the reliability and validity of behavioral instruments currently in use to select and train MWDs for their working life.

Acknowledgements

This work was funded by an NSF Award 0731216 to SDG. Timothy Bartlett, James Dalton, Lacy Smith, and Lisa Maze provided essential logistical support, and we thank other members of 341 TRS for caring for and training MWDs at Lackland. Two anonymous reviewers and the AABS Editorial Board provided valuable comments on an earlier version.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.applanim.2010.08.007.

References

Beerda, B., Schilder, M.B.H., van Hooff, J.A.R.A.M., de Vries, H.W., 1997. Manifestations of chronic and acute stress in dogs. Appl. Anim. Behav. Sci. 52, 307–319.
Bell, A.M., Hankison, S.J., Laskowski, K.L., 2009. The repeatability of behaviour: a meta-analysis. Anim. Behav. 77, 771–783.
Benus, R.F., Koolhaas, J.M., van Oortmerssen, G.A., 1987. Individual differences in behavioural reaction to a changing environment in mice and rats. Behaviour 100, 105–122.
Cattell, R.B., 1966. The scree test for the number of factors. Sociol. Meth. Res. 1, 245–276.
Champness, K.A., 1996. Development of a Breeding Program for Drug Detector Dogs: Based on Studies of a Breeding Population of Guide Dogs. Department of Agriculture and Resource Management, University of Melbourne, Melbourne, Australia.
Draper, T.W., 1995. Canine analogs of human personality factors. J. Gen. Psychol. 122, 241–252.
Field, A., 2005. Discovering Statistics using SPSS, second ed. Sage Publications, London.
Fuchs, T., Gaillard, C., Gebhardt-Henrich, S., Ruefenacht, S., Steiger, A., 2005. External factors and reproducibility of the behaviour test in German shepherd dogs in Switzerland. Appl. Anim. Behav. Sci. 94, 287–301.
Goddard, M.E., Beilharz, R.G., 1982. Genetic and environmental factors affecting the suitability of dogs as guide dogs for the blind. Theor. Appl. Genet. 62, 97–102.
Goddard, M.E., Beilharz, R.G., 1983. Genetics of traits which determine the suitability of dogs as guide-dogs for the blind. Appl. Anim. Ethol. 9, 299–315.
Goddard, M.E., Beilharz, R.G., 1984. A factor analysis of fearfulness in potential guide dogs. Appl. Anim. Behav. Sci. 12, 253–265.
Goddard, M.E., Beilharz, R.G., 1985. Individual variation in agonistic behaviour in dogs. Anim. Behav. 33, 1338–1342.
Goddard, M.E., Beilharz, R.G., 1986. Early prediction of adult behavior in potential guide dogs. Appl. Anim. Behav. Sci. 15, 247–260.
Goldberg, L.R., Digman, J.M., 1994. Revealing the structure in the data: Principles of exploratory factor analysis. In: Strack, S., Lorr, M. (Eds.), Differentiating Normal and Abnormal Personality. Springer, New York, pp. 216–242.
Gorsuch, R.L., 1983. Factor Analysis, second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
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., Kwan, V.S.Y., John, O.P., 2003. A dog's got personality: a cross-species comparative approach to evaluating personality judgements in dogs and humans. J. Pers. Soc. Psychol. 85, 1161–1169.
Graham, L.T., Gosling, S.D., 2009. Temperament and personality in working dogs. In: Helton, W.S. (Ed.), Canine Ergonomics: The Science of Working Dogs. CRC Press, New York, pp. 63–81.
Haverbeke, A., De Smet, A., Depiereux, E., Giffroy, J.M., Diederich, C., 2009. Assessing undesired aggression in military working dogs. Appl. Anim. Behav. Sci. 117, 55–62.
Helton, W.S., 2009. Canine Ergonomics: the Science of Working Dogs. CRC Press, New York.
Hennessy, M.B., Voith, V.L., Mazzei, S.J., Buttram, J., Miller, D.D., Linden, F., 2001. Behavior and cortisol levels of dogs in a public animal shelter, and an exploration of the ability of these measures to predict problem behavior after adoption. Appl. Anim. Behav. Sci. 73, 217–233.
John, O.P., Robins, R.W., 1993. Determinants of interjudge agreement on personality traits: the big five domains, observability, evaluativeness, and the unique perspective of the self. J. Pers. 61, 521.

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.

Kaminski, J., Bräuer, J., Call, J., Tomasello, M., 2009. Domestic dogs are sensitive to a human's perspective. Behaviour 146, 979–998.

Keiichi, T., Tucker, D., 1985. Responsiveness of the olfactory receptor cells in dog to some odors. Comp. Biochem. Phys. A 81, 7–13.

Koolhaas, J.M., Korte, S.M., De Boer, S.F., Van Der Vegt, B.J., Van Reenen, C.G., Hopster, H., De Jong, I.C., Ruis, M.A.W., Blokhuis, H.J., 1999. Coping styles in animals: current status in behavior and stress-physiology. Neuro. Biobehav. Rev. 23, 925–935.

Ley, J., Bennett, P., Coleman, G., 2008. Personality dimensions that emerge in companion canines. Appl. Anim. Behav. Sci. 110, 305–317.

Ley, J.M., McGreevy, P., Bennett, P.C., 2009. Inter-rater and test–retest reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R). Appl. Anim. Behav. Sci. 119, 85–90.

Lit, L., 2009. Evaluating learning tasks commonly applied in detection dog training. In: Helton, W.S. (Ed.), Canine Ergonomics: The Science of Working Dogs. CRC Press, London, pp. 99–114.

Lorenzo-Seva, U., ten Berge, J.M.F., 2006. Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology 2, 57–64.

Maejima, M., Inoue-Murayama, M., Tonosaki, K., Matsuura, N., Kato, S., Saito, Y., Weiss, A., Murayama, Y., Ito, S.I., 2007. Traits and genotypes may predict the successful training of drug detection dogs. Appl. Anim. Behav. Sci. 107, 287–298.

McCrae, R.R., Zonderman, A.B., Bond, M.H., Costa, P.T., Paunonen, S.V., 1996. Evaluating replicability of factors in the revised NEO personality inventory: confirmatory factor analysis versus procrustes rotation. Am. Psych. Assoc., 552–566.

Moulton, D.G., Marshall, D.A., 1979. Quantitative analysis of nasal air-flow in the dog during sniffing. Am. Zoologist 19, 864.

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

Rooney, N.J., Bradshaw, J.W.S., Almey, H., 2004. Attributes of specialist search dogs—a questionnaire survey of UK dog handlers and trainers. J. Forensic Sci. 49, 300–306.

Rooney, N.J., Gaines, S.A., Bradshaw, J.W.S., Penman, S., 2007. Validation of a method for assessing the ability of trainee specialist search dogs. Appl. Anim. Behav. Sci. 103, 90–104.

Rooney, N., Gaines, S., Hiby, E., 2009. A practitioner's guide to working dog welfare. J. Vet. Behav.: Clin. Appl. Res. 4, 127–134.

Ruefenacht, S., Gebhardt-Henrich, S., Miyake, T., Gaillard, C., 2002. A behaviour test on German Shepherd dogs: heritability of seven different traits. Appl. Anim. Behav. Sci. 79, 113–132.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Slabbert, J.M., Odendaal, J.S.J., 1999. Early prediction of adult police dog efficiency—a longitudinal study. Appl. Anim. Behav. Sci. 64, 269–288.

Svartberg, K., 2002. Shyness-boldness predicts performance in working dogs. Appl. Anim. Behav. Sci. 79, 157–174.

Svartberg, K., 2005. A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Appl. Anim. Behav. Sci. 91, 103–128.

Svartberg, K., Forkman, B., 2002. Personality traits in the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci. 79, 133–155.

Svartberg, K., Tapper, I., Temrin, H., Radesäter, T., Thorman, S., 2005. Consistency of personality traits in dogs. Anim. Behav. 69, 283–291.

Tabachnick, B.G., Fidell, L.S., 1996. Using Multivariate Statistics. Harper-Collins, New York.

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.

Wells, D.L., 2009. Sensory stimulation as environmental enrichment for captive animals: a review. Appl. Anim. Behav. Sci. 118, 1–11.

Wilsson, E., Sundgren, P.-E., 1997a. The use of a behaviour test for selection of dogs for service and breeding. II. Heritability for tested parameters and effect of selection based on service dog characteristics. Appl. Anim. Behav. Sci. 54, 235–241.

Wilsson, E., Sundgren, P.-E., 1997b. The use of a behaviour test for the selection of dogs for service and breeding. I: Method of testing and evaluating test results in the adult dog, demands on different kinds of service dogs, sex and breed differences. Appl. Anim. Behav. Sci. 53, 279–295.

Wilsson, E., Sundgren, P.-E., 1998. Behaviour test for eight-week old puppies—heritabilities of tested behaviour traits and its correspondence to later behaviour. Appl. Anim. Behav. Sci. 58, 151–162.

Zwick, W.R., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. Psychol. Bull. 99, 432–442.