

## Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program

Erik Wilsson<sup>a,b,\*</sup>, David L. Sinn<sup>c,d</sup>

<sup>a</sup> Swedish Armed Forces Dog Training Center, Box 194, SE-195 24 Märsta, Sweden

<sup>b</sup> S Tisslegården 3, SE 542 95 Mariestad, Sweden

<sup>c</sup> University of Texas, Department of Psychology, 1 University Station, A8000, Austin, TX 78712, USA

<sup>d</sup> University of Tasmania, School of Zoology, Private Bag 5, Hobart, TAS 7001, Australia

### ARTICLE INFO

#### Article history:

Accepted 27 August 2012

Available online 15 September 2012

#### Keywords:

Behavior test  
Behavioral ratings  
Subjective ratings  
Working dogs  
Predictive validity

### ABSTRACT

Consistent behavioral variation within and between individuals is ubiquitous in all working dog populations. Most working dog programs have recognized this fact, and have subsequently attempted to quantify behavior through the use of standardized tests. Standardized tests may employ several measurement methods, but two common ones are behavioral ratings and subjective ratings. The former is characterized by a rating for a behavior (e.g., reaction to a noise) usually based on a single observation or test; the latter is characterized by a rating for a trait (e.g., confidence) that is based across multiple observations of behavior. The main difference between the two rating methods is the level of aggregation, or intuition, that is required by the human observer. Measurement theory predicts that ratings based on multiple observations (i.e., subjective ratings) should be more reliable because measurement error is reduced. However, subjective ratings, by definition, may be susceptible to observer bias, in which case ratings based on fewer, but better-defined observations (i.e., behavioral ratings) could result in greater reliability. In either case, the ultimate criterion of most working dog programs is the predictive validity of measured behaviors in standardized tests. To the best of our knowledge, the relative predictive validity of subjective and behavioral ratings has yet to be tested within the same working dog population. Here we analyzed behavioral test results along with training outcomes from a large sample (~400) of German shepherd dogs, aged 15–18 months, all bred at the Swedish Armed Forces breeding kennel. Behaviors observed in the test were measured using 25 behavioral and 13 subjective ratings. Data reduction and confirmatory techniques identified five underlying dimensions in behavioral ratings (confidence, physical engagement, social engagement, aggression, and environmental sureness) and three in subjective ratings (engagement, confidence, and aggression). Both ratings methods correctly classified a high percentage of dogs that did/did not complete training (70.3–78.3%). However, only minor differences in predictive validity were observed between the two measurement methods (1.7–6.6%). Engagement and confidence, irrespective of measurement method, were the strongest predictors of training completion, but the two ratings methods identified different aspects of engagement and confidence that may be important to training outcomes in this and other working dog programs. Taken

\* Corresponding author at: Swedish Armed Forces Dog Training Center, Box 194, SE-195 24 Märsta, Sweden. Fax: +46 8 597 888 08.  
E-mail address: [tisslegarden@telia.com](mailto:tisslegarden@telia.com) (E. Wilsson).

together, our results suggest that in some cases, the use of subjective versus behavioral ratings may be inconsequential from the standpoint of prediction to training criterion. Further empirical verification is needed, along with improvements in the explicit definition and measurement of 'success' in this and other working dog programs.

© 2012 Elsevier B.V. All rights reserved.

## 1. Introduction

Behavioral variation is ubiquitous in all dog populations (Diederich and Giffroy, 2006; Jones and Gosling, 2005; Svartberg, 2007), including those that have been selected, bred, and raised for a particular job that requires particular behaviors (Graham and Gosling, 2009; Willis, 1995). In other words, idiosyncratic, individual differences in dog behavior are always observed, even when they are selected against, and a dog's personality often impacts its success in working dog programs (Goddard and Beilharz, 1986; Maejima et al., 2007; Saetre et al., 2006; Sinn et al., 2010; Slabbert and Odendaal, 1999; van der Waaij et al., 2008; Wilsson and Sundgren, 1997a, 1997b).

Here we use the term 'personality' synonymously with an individual dog's general disposition. More specifically, personality can be defined as 'correlated suites of behaviors', whereby the correlations between observed behaviors occur through time, across different functional contexts, or some combination of the two (Gosling, 2001; Sih et al., 2004). For example, a dog's 'aggressiveness', indexed by its body postures, piloerection and baring of teeth may be correlated with its 'boldness', indexed by its confidence in novel or noisy conditions, and its willingness to explore novel objects. Similarly, correlations through time occur, for example, when an individual's boldness or aggression at one time or life stage is correlated with its boldness or aggression at another.

Personality variation, historically, is well recognized in most working dog programs (e.g., Humphrey, 1934; Martínek et al., 1971; Pfaffenberger et al., 1976); as such, standardized behavioral testing that attempts to measure dog personality is common and is often used to select dogs for training, breeding, or for procurement. In standardized tests, a sample of dogs is subjected to the same/similar stimuli, and human observers generally attempt to measure a large number of observed behaviors.

These multiple observed behaviors can be measured using several different techniques, but the most common methods are through the use of either codings or ratings (Jones and Gosling, 2005; Vazire et al., 2007). Behavioral coding methods attempt to narrowly define discrete, observable behaviors (e.g., barking, yawning, or tail wagging) and then use frequency counts and durations of those observed behaviors to code subject behavior (e.g., Netto and Planta, 1997). Behavior codings are often based on species or population ethograms, and are considered to be relatively objective in comparison to ratings methods (but see Vazire et al., 2007). Ratings methods, on the other hand, rely on human observers to intuitively aggregate and interpret dog behavior for particular, pre-defined behavioral traits. Ratings are usually based on Likert scales, for example, from one to five or one to seven, where the

number reflects the severity of the behavior displayed by the subject animal as deemed by a human observer. Two types of ratings methods can be distinguished (Svartberg, 2007). The first, termed 'behavioral ratings (BR)', are ratings that are based on observed behaviors in a particular test situation, and attempt to rate behavior as objectively as possible. For example, Svartberg (2002) used 23 different a priori BR scales that were assessed in nine different standardized test situations to characterize dog personality in German shepherd and Belgian Tervuren working dogs. In one of these subtests, dogs were given a single rating of their 'sociability', based on their avoidance (lower scores) or approach (higher scores) toward a stranger that approached and physically greeted them.

The second type of rating method is termed 'subjective ratings' (Block, 1961). Subjective ratings (SR) attempt to evaluate global behavior of the subject, and are based on a human scorer's overall perception of the dog's aggregate behavioral disposition, sometimes (but not always) irrespective of a test situation (e.g., Hsu and Serpell, 2003; Ley et al., 2009). Subjective ratings are usually best used when human scorers know their subjects well (e.g., Gosling, 1998), and can be used to evaluate global behavior of subjects across many different test scenarios (as opposed to making a behavioral rating that is based on only a single, specific test instance).

BR's, with less reliance on observer intuition to aggregate observed behaviors, are generally considered to be less subjective than SR methods, and are, as such, anecdotally considered to give more reliable information. However, both SR (Ruefenacht et al., 2002; Wilsson and Sundgren, 1997a) and BR (Svartberg and Forkman, 2002) reliability has been evaluated empirically, and results indicate that both ratings methods often correlate with behavioral codings made on the same subjects (De Meester et al., 2008; Hewson et al., 1998; van den Berg et al., 2006). What is less well known is whether different measurement methods have different degrees of predictive validity with regards to working outcomes (e.g., certification in a training program). Different rating methods used within the same working program may differ in their predictive validity depending on how closely the BR or SR relates to the outcome of interest (usually an aspect of success), which is often a narrowly defined dichotomous measure (e.g., 'success or failure'). Theoretically, when reliability of both BR and SR methods are equal (e.g., both provide generally applicable information equally well), SR methods may have greater predictive validity if the behaviors relevant to the outcome measurement are not captured by the more restrictive BR methods. On the other hand, if outcome measurements are adequately captured by BR methods, we may expect BR to have greater or similar levels of predictive validity relative to SR methods. Likewise, if neither

ratings method captures relevant information regarding the traits that are important to success in training or to actual work, then predictive validity of both methods would be low. To the best of our knowledge, no dog studies have previously assessed both BR and SR methods simultaneously to determine the relative predictive validity of each (for an example from the primate literature, see [Uher and Asendorpf, 2008](#)). Clearly, it is important for a working dog program to know whether a relatively more objective measurement method (e.g., BR) or a measurement method that includes greater levels of observer aggregation (e.g., SR) is superior in capturing variance between dogs in their ability to be successful workers.

The breeding program of the Swedish armed forces (SAF) has the aim of improving behavioral traits in working dogs that are asked to perform substance detection and personnel protection duties. In the SAF program, dogs are selected for training based on observed behavior in a standardized behavioral test. The behavioral test consists of 14 different sub-tests; at least one BR is given in each sub-test and a number of SR's are given based on global test behavior. Individual dogs are tracked through training, and subsequent success criteria (successful completion of training) are recorded for each individual, allowing for a test of the relative predictive validity of the BR's and SR's that are given during the 'pre-screening' SAF standardized behavior test. In order to test the relative predictive validity of the two measurement methods, we first examined the inter-correlations of the BR's and SR's ourselves (i.e., identified underlying dimensions of behavior, or personality traits via principal components analysis) to reduce the number of subsequent variables needed for predictive purposes. Because predictive validity of the two methods was similar (see Section 3.2), we also report correlations between personality trait scores identified by BR and SR methods to identify aspects of the two measurement methods that may impact the ability of working dog programs to measure relevant behaviors.

## 2. Materials and Methods

### 2.1. Subjects

The SAF breeding program began in 2005 in order to supply the SAF with high quality detection and patrol dogs. The SAF program currently aims each year to produce at least 50 high quality working dogs per 200 puppies produced per annum. The original breeding bitches in the SAF program came from various sources across Europe while most of the studs were already active working dogs in Sweden, mainly police dogs or patrol dogs. Since the start of 2011, all new breeders, both males and females, have been recruited exclusively from puppies produced within the SAF program itself.

At birth, puppies are group-housed with their litter and mother until weaning at 8 weeks of age, when they are fostered to members of the Swedish public who volunteer as puppy raisers. Local SAF instructors provide assistance during puppy rearing by providing advice and

support to volunteers upon request. Puppy raisers are asked to do ordinary house training and some basic obedience training, but raisers are also asked not to do any other specialized training beyond this. Between the age of 15–18 months, dogs are returned to the SAF program and given a behavioral test (Section 2.2) at one of five different locations. Initial acceptance into the SAF training program is determined by a single test leader's (hereafter, 'TL') subjective opinion formed during the entire behavioral test procedure. Dogs that 'pass' this behavioral test are immediately placed into training, normally for 8–10 months. If dogs attain training certification, they enter their working life within the SAF, local police forces, Swedish customs, or other government authorities. Dogs not passing the initial behavioral test, or dogs being rejected during the 8–10 month SAF training period, are sold as companion animals, usually to the host family previously having cared for the dog. Rejections from SAF training are most normally for behavioral reasons (i.e., high levels of chronic stress in the kennel environment, and/or not capable of completing training tasks) but sometimes occur if medical conditions, not detected earlier in life, develop. Decisions to certify a dog/reject a dog from training are made by multiple SAF training staff, not including the TL.

The sample we present here consisted initially of 496 German shepherd dogs that were bred from 91 litters at the SAF kennel between the years 2005 and 2009. Only dogs with the full complement of ratings and those that were deemed medically fit were used in statistical analyses here; thus sample sizes vary for each analyses and are given below (Section 2.5).

### 2.2. General behavior test procedure

The standardized behavior test is similar to that used in other working dog programs (e.g., [Svartberg and Forkman, 2002](#); [Wilsson and Sundgren, 1997a](#)). Behavior tests were always performed during the day between 08:00 and 16:00 h. Subject dogs were always accompanied by the puppy raiser, and all dogs were always assessed by the same TL (i.e., a single test leader gave all BRs and SRs in our dataset). All dogs received the full complement of 12 sub-tests in the same day; all sub-tests combined take on average 40 min to complete. In 36 cases where the subject dog became extremely fearful, without being able to calm down, testing was immediately halted.

Behavioral tests were performed in the same order at each test time for each dog (Section 2.3). Order of selecting dogs throughout the day was haphazard, but males were tested before females in order to avoid distraction from pre-estrus females on males. Except for two sub-tests (visual startle response and gunfire, Sections 2.3.9 and 2.3.12, respectively), sub-tests were carried out indoors in a large garage-like building with concrete floor and multiple, smaller attached rooms. In each sub-test one or more BR's were given ([Table 1](#)). After the full complement of sub-tests was completed the TL gave 13 additional SR's ([Table 2](#)). All ratings were given on a scale from 1 to 5, unless otherwise noted.

**Table 1**

Operational definitions of behavior ratings given during the 12 sub-tests of the standardized behavior test. TL = test leader.

| Sub-test                | Rating                | 1                                                              | 2                                                        | 3                                                                               | 4                                                                                | 5                                                                           |
|-------------------------|-----------------------|----------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------------------------|----------------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Affability and handling | Affability            | Rejects contact, withdraws.                                    | Does not reject contact. No withdrawal.                  | Does not reject contact. No withdrawal. Makes contact within 15 s.              | Makes spontaneous contact without jumping and vocalization.                      | Intensive contact with vocalization or jumping at person.                   |
| Affability and handling | Handling              | Rejects, growls, tries to bite/escape.                         | Pulls away, seeks support from handler.                  | Accepts handling.                                                               | Accepts handling, seeks contact with TL.                                         | Overwhelming contact toward TL when handled.                                |
| Leash                   | Leash                 | Acts on its own. No contact with handler.                      | Acts on its own. Attentive when handler demanding.       | Follows without handler demanding.                                              | Dependent. Looking for confirmation from handler.                                |                                                                             |
| Tug-of-war              | Tug-of-war            | Does not take rag.                                             | Takes rag. Let's go before handler pulls.                | Takes rag. Let's go when handler pulls.                                         | Pulls hard on rag but let's go when handler pulls back hard or make loud noises. | Pulls hard on rag. Does not let go despite hard resistance or disturbances. |
| Retrieving              | Chasing               | Does not run after ball.                                       | Starts running but stops before reaching the ball.       | Runs after and takes the ball.                                                  | Runs after and carries ball back to handler.                                     | Runs after with high intensity and carries ball back to handler.            |
| Retrieving              | Interest in object    | Does not take ball.                                            | Grabs ball but let's go immediately.                     | Grabs ball, carries less than 5 s.                                              | Grabs ball, carries.                                                             | Grabs ball intensely, carries.                                              |
| Dark room               | Reaction in dark room | Attempts to leave the room.                                    | Freezes when light is out.                               | Walks into the room less than 3 m and stops.                                    | Investigates but is affected. Goes to puppy raiser but with help.                | Investigates without hesitation or finds puppy raiser directly.             |
| Metal stair             | Metal stair           | Refuses to follow puppy raiser.                                | Starts to follow but then refuses after some steps.      | Manages to walk the whole stairway but with major hesitation.                   | Easily walks the stairway, but with slight hesitation.                           | Walks the stairway without hesitation.                                      |
| Unstable table          | Reaction on table     | Tries to escape before table is moving.                        | Tries to escape when table is moving.                    | Visibly uncomfortable with a low tense body posture but does not try to escape. | Tense but not annoyed.                                                           | Unaffected. Dog is just standing still while table is moving.               |
| Unstable table          | Object                | Does not take ball.                                            | Takes ball but let's go when table moves.                | Holds ball despite moving table.                                                |                                                                                  |                                                                             |
| Acoustic startle        | Flight distance       | Escapes <sup>a</sup> >5 m.                                     | Escapes 2–5 m.                                           | Escapes 1–2 m                                                                   | Takes one or two steps backwards without escaping.                               | No fear response.                                                           |
| Acoustic startle        | Secondary response    | Does not want to investigate buckets despite encouragement.    | Investigates buckets but only with major encouragement.  | Investigates buckets with some help from handler.                               | Investigates buckets without help but with hesitation.                           | Investigates buckets without hesitation.                                    |
| Acoustic startle        | Lasting effect        | Very affected. Persistent avoidance.                           | Affected. Is passing but active avoidance.               | A bit affected. Change in body posture when passing.                            | Unaffected but is looking at object.                                             | No lasting effect.                                                          |
| Visual startle          | Flight distance       | Escapes <sup>a</sup> >5 m.                                     | Escapes 2–5 m.                                           | Escapes 1–2 m.                                                                  | Jerks without escaping.                                                          | No fear response.                                                           |
| Visual startle          | Aggression            | No sign of aggression, i.e. piloerection, barking or growling. | Some sign of aggression. Barking or piloerection.        | Clear signs of aggression, piloerection or growling.                            | Strong aggression, including mouth threat.                                       | Attacks and bites the coverall.                                             |
| Visual startle          | Secondary response    | Does not investigate coveralls.                                | Investigates coveralls with major encouragement.         | Investigates coveralls with some encouragement.                                 | Investigates coveralls without help but with hesitation.                         | Investigates coveralls without hesitation.                                  |
| Visual startle          | Lasting effect        | Very affected. Persistent avoidance of coverall.               | Affected. Passes but shows active avoidance of coverall. | Mildly affected. Change in body posture when passing.                           | Unaffected but is looking at object when passing.                                | No lasting effect.                                                          |
| Gradual visual startle  | Fearfulness           | Freezes or tries to escape.                                    | No reaction.                                             | Switches between fight/flight responses.                                        | Strong fight reaction toward figure, but without attack.                         | Attacks and bites figure.                                                   |
| Gradual visual startle  | Aggression            | No sign of aggression, i.e. piloerection, barking, growling.   | Some sign of aggression. Barking or piloerection.        | Clear sign of aggression, piloerection.                                         | Strong aggression, growls, mouth threat.                                         | Bites.                                                                      |
| Gradual visual startle  | Secondary response    | Does not investigate figure.                                   | Investigates figure with major encouragement.            | Investigates figure with some encouragement.                                    | Investigates figure without help but with hesitation.                            | Investigates figure without hesitation.                                     |

|                        |                                                      |                                                       |                                                       |                                            |
|------------------------|------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------|--------------------------------------------|
| Gradual visual startle | Very affected. Persistent avoidance of figure.       | Mildly affected. Change in body posture when passing. | Unaffected but is looking at object when passing.     | No lasting effect.                         |
| Search                 | Does not search.                                     | Searches but stops.                                   | Searches intensively, uses mouth or paws.             | Searches intensively, uses mouth and paws. |
| Search                 | Does not start search.                               | Searches <10 s.                                       | Searches <2 min.                                      | Searches <2 min.                           |
| Gunfire                | Very fearful, tries to escape <sup>a</sup> in leash. | Stops playing, not playful afterwards.                | No reaction.                                          | Very excited, pulling on leash             |
| Gunfire                | No reaction.                                         | Stops playing, looks in direction of gunfire.         | Stops playing, want to investigate, whines, whimpers. | in direction of gunfire cannot be calmed.  |

<sup>a</sup> Escapes means turning away 180° from stimulus source and moving away.

**Table 2**

Operational definitions of subjective ratings given during the standardized behavioral test. Each trait is rated from 1 to 5, with 1 representing 'low expression' of a trait, and 5 representing 'high expression' of a trait.

| Trait                               | Definition                                                                                                                                                                                                                                                                                        |
|-------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Affability <sup>a</sup>             | The dog's comfort level and interaction with people. Synonymous with 'sociability' with humans.                                                                                                                                                                                                   |
| Competitiveness                     | Displaying a strong desire to have sole possession of objects.                                                                                                                                                                                                                                    |
| Hunting drive <sup>a</sup>          | The dog's willingness, vigor, or enthusiasm to run after a moving object.                                                                                                                                                                                                                         |
| Environmental sureness <sup>a</sup> | The dog's ability to cope with a variety of noxious physical environmental stimuli or disturbances.                                                                                                                                                                                               |
| Courage <sup>a</sup>                | The absence of fearful behavior toward real or imagined danger; such as the ability to rebound from unnerving situations.                                                                                                                                                                         |
| Nerve stability <sup>b</sup>        | The appropriateness of the dog's reaction to a certain situation. This includes the dog's ability to adapt to various types of non-fearful situations, to concentrate when highly aroused or in a situation of conflict, as well as its ability to relax and to overcome a frightening situation. |
| Hardness <sup>a</sup>               | A mental and/or physical resiliency to unpleasant experiences. Hard dogs are highly "recoverable" after disturbances.                                                                                                                                                                             |
| Liveliness                          | The dog's general mental and physical arousal.                                                                                                                                                                                                                                                    |
| Sharpness <sup>a</sup>              | An act of aggression or agonistic interaction. It can be appropriate or inappropriate, and involve a threat, challenge or contest.                                                                                                                                                                |
| Defense drive <sup>b</sup>          | The tendency for the dog to defend itself or its handler. In most cases defense is combined with aggression. However a dog may show defensive tendencies without being aggressive.                                                                                                                |
| Co-operation <sup>b</sup>           | The tendency to be influenced by the handler without being given a direct command or sign.                                                                                                                                                                                                        |
| Prey drive                          | The dog's interest in objects, its willingness to search for, to bite and to carry them in the mouth.                                                                                                                                                                                             |
| Curiosity                           | A tendency to explore and to investigate new things.                                                                                                                                                                                                                                              |

<sup>a</sup> Definition from SWGDOG (2012).

<sup>b</sup> Definition from Wilsson and Sundgren (1997a).

## 2.3. Specific sub-test procedure and behavioral ratings

### 2.3.1. Affability and handling sub-test

The dog on leash is led through a group of 2–5 passive persons not familiar to the dog for approximately 1 min. The passive persons are instructed not to actively interact or make contact with the dog. Next, the TL takes over the leash, and then physically examines the dog (i.e. teeth inspection, palpation of legs and paws) for approximately 1 min. Two BRs are given during this sub-test, one based on the subject's reaction to the unknown people (affability), and one based on the subject's reaction to being handled by the TL (handling). For both ratings, higher scores indicated dogs that were more sociable and less withdrawn,

fearful, or aggressive, and lower scores indicated the opposite.

### 2.3.2. Leash sub-test

Next, the TL walks the dog on leash for 1 min while haphazardly and repeatedly changing direction. One BR is given during this sub-test, based on the subject's attentiveness to the test leader (leash). Higher scores indicated more dependence on cues given by the TL by dogs while on lead, and lower scores were given to dogs that paid less attention to the TL while walking. Leash response was rated on a 1–4 scale.

### 2.3.3. Tug-of-war sub-test

After stopping from the leash sub-test, the TL invites the dog to a tug-of-war with a cotton rag. The dog is encouraged to bite and to pull on the rag for 2 min. One BR is given, based on the subject's interest in playing tug-of-war. Higher scores were given to dogs with greater interest.

### 2.3.4. Retrieving sub-test

Next, the TL removes the cotton rag from visual sight, and rolls a tennis ball over the floor approximately 15 m away; the dog is allowed to chase and to retrieve the ball. Chase and retrieval is performed three times. Based on observations across all three episodes of retrieving, two BR's are given. The first is based on the subject's intensity of chasing the ball (chasing), and the second BR is based on the subject's intensity of physical possession of the ball (interest in object). Higher chasing scores were given to dogs that were faster and more enthusiastic while chasing, and higher interest in object scores were given to those individuals that were physically possessive of the ball once retrieved.

### 2.3.5. Dark room sub-test

After the retrieving test, the TL leads the subject on leash to the door of a small, 5 m × 5 m windowless room. The puppy raiser enters the room in view of the subject, but then crouches in an opposite corner, behind an opaque barrier where the subject is unable to see. The light is then turned off and the puppy raiser calls for the subject. One BR is given in this sub-test (reaction in a dark room). Higher scores are given to dogs that investigate the room, and go to the puppy raiser with less encouragement. Lower scores are given to dogs that freeze or try to withdraw from the dark room.

### 2.3.6. Metal stair sub-test

Next, the TL takes the leash and follows the puppy raiser who walks outside of the dark room into the main test building area to a flight of steep metal stairs. The TL and subject on leash follow the puppy raiser up and down the stairs. After the subject goes up and down the steep stairs on lead, the situation is repeated twice more. One BR is given (metal stair) based on only the third response to the metal stairs. Higher scores for metal stair are given to dogs that confidently walk up and down the stairs, while lower scores are given to dogs that hesitate or that refuse to follow the puppy raiser.

### 2.3.7. Unstable table sub-test

After the metal stairs, the subject is led on leash by the puppy raiser to an adjacent room and is asked to jump up on a table approximately 60 cm high. Once the subject has settled on the top of the table the TL wobbles the table back and forth, displacing it 2–3 cm. While the table is still moving the TL offers a tennis ball to the subject. Two BRs are given during this sub-test, based on a subject's reaction to the moving table (reaction on table) and to the offered ball (object). High reaction on table scores were given to dogs that were more confident, and made no attempts to escape to the floor; lower scores were given to dogs that were noticeably affected by standing on the table or having it move, or that tried to escape. High object scores were given to dogs that took and held on to the tennis ball during table movement; lower object scores were given to dogs that did not take the ball or let go of the ball when the table moved.

### 2.3.8. Acoustic startle sub-test

Next, while the subject is being walked on leash by the puppy raiser to a different room in the test building, a pair of steel buckets is dropped on the floor approximately 2 m from the subject. The puppy raiser is told to let go of the leash as the buckets are dropped. The dog is encouraged to investigate the buckets; the puppy raiser then regains the leash and walks the subject around the room, passing the fallen buckets three more times. Three BR's were taken in this sub-test (flight distance, secondary response, and lasting effect). Higher scores for flight distance were given to dogs that gave no fear response or did not react by trying to escape in response to the falling buckets. Lower flight distance scores were given to dogs that tried to escape, i.e., turned around 180° from the stimulus source and initially increased their distance from the disturbance. Higher scores for secondary response were given to dogs that immediately investigated the buckets, and lower scores were given to dogs that did not investigate the buckets afterwards. Low lasting effect scores were given to dogs that were visibly disturbed, i.e., attempted to avoid the buckets when passing them on a leash. High scores were given to dogs that did not pay attention to the buckets even when passing by on the leash the first time.

### 2.3.9. Visual startle sub-test

After the acoustic startle sub-test, the puppy raiser leads the dog away from the area of the buckets. Two meters in front of the dog a pair of life-sized coveralls, resulting in the shape of an 'X', is pulled into the air by the TL, and the puppy raiser releases the leash. After the dog has made its initial reaction and investigated the doll, the puppy raiser regains the leash and walks the dog past the coveralls until no sign of attention is observed. Four BR's are given during this sub-test (flight distance, aggression, secondary response, and lasting effect). Flight distance has the identical definition of the flight distance BR given in acoustic startle sub-tests. Higher aggression scores are given to dogs that growl, threaten, and bite the coveralls, while lower aggression scores are given to dogs that show very little or no signs of threatening postures or vocalizations. Higher scores for secondary response are given to dogs that immediately investigated the coverall and lower scores were given to

dogs that refused to investigate or did so only after being encouraged by the handler. Lower lasting effect scores are given to dogs that look at or continually attempt to physically avoid the coveralls. Higher lasting effect scores are given to dogs that pay no visible attention to the coveralls when passing it on leash.

#### 2.3.10. Gradual visual startle sub-test

A paper figure, the top half of a person, is mounted on two wooden planks. The eyes of the figure are prominent and the planks are held at an angle so that it is facing the subject. The puppy raiser, standing beside the TL, holds the subject on leash while the figure/planks are pulled into view 15 m from the subject. The figure/planks are then slowly moved toward the subject and stopped at 3 m distance from the subject, and the subject is released from the lead. Once the subject has physically inspected the paper figure the puppy walker takes the subject on leash and walks back and forth close to the figure three times. Four BR's are given during this sub-test (fearfulness, aggression, secondary response, and lasting effect). Higher fearfulness scores are given to dogs that do not attempt to escape, and even proactively approach or attack the figure; lower fearfulness scores are given to dogs that have aversive reactions, or those that freeze or try to escape. Higher aggression scores are given to dogs that physically threaten or bite the figure, low scores are given to dogs showing no signs of aggression, i.e., barking, piloerection, etc. Higher secondary response scores are given to dogs that immediately investigated the figure once they were sent free, lower scores are given to dogs not willing to investigate the figure even when helped by the puppy raiser. Higher lasting effect scores are given to dogs that paid no attention to the figure whatsoever even when passing it close on leash the first time. Lower scores are given to dogs that continue to look at or try to avoid the figure even after passing it the third time.

#### 2.3.11. Search sub-test

The puppy raiser next leads the subject away from the paper figure by its collar, and then rolls a tennis ball toward the TL who hides the ball under a wooden pallet. When the ball has been hidden the subject is released and allowed to independently search. If the subject does not find the ball within 2 min or if it loses interest after this time, it is encouraged and helped by the TL until it retrieves the ball. Hiding and searching for the tennis ball is repeated three times. Two BR's are given in this sub-test (intensity and persistence). Higher scores for intensity are given to dogs that show greater interest in searching and tended to use different search strategies, e.g., mouthing the pallet and/or using paws, while lower intensity scores are given to dogs that did not show interest in searching. Higher scores for persistence were given to dogs that were willing to spend more time unassisted searching for the hidden tennis ball, while lower persistence scores are given to dogs that spent less time searching independently.

#### 2.3.12. Gunfire sub-test

In the final sub-test, the subject is taken outdoors on leash by the puppy raiser. Two, 9-mm blank shots are fired

from 25 m away out of the visual sight of the subject. The TL then engages the subject in a tug-of-war for approximately 1 min and another two shots are fired. Two BRs are given during the gunfire sub-test (fearfulness and curiosity). Fear BR's are given on a 1–4 scale. Higher fearfulness scores are given to dogs that are passive or do not react to the gunfire; lower fearfulness scores are given to dogs that quit playing or that try to escape after gunfire. Higher curiosity scores are given to dogs that show a desire to investigate, whine or whimper, or pull toward the place where the gunfire came from, while lower curiosity scores are given to dogs that simply stop playing but do not show an interest in investigating.

### 2.4. Subjective ratings

Thirteen SR's were given by the TL to each subject at the completion of all sub-tests; each SR was given on a 1–5 scale with 1 indicating a lack of expression and a 5 indicating a high degree of expression of that trait (Table 2). All SRs were based on observations across multiple sub-tests; two SRs (i.e., 'liveliness' and 'curiosity') were based on observations across all sub-tests. As opposed to the BR ratings, which were not available to anyone besides the TL, SR ratings were available to SAF training staff. However, in practice, SRs are rarely consulted by SAF trainers; dogs have already been selected for training and represent a high economical value, so all efforts are made by trainers in order to have the dog succeed in training.

### 2.5. Data analysis

After removal of dogs that were identified as having medical issues, the total *N* for dogs with complete behavioral test records was 383. Given the sample size and exploratory nature of our study, for all analyses we first randomly chose 70% of the subjects to serve as our training data set (i.e., data from which model parameters were estimated) and 30% of subjects as our test data set (i.e., data from which fitted parameters were tested to predict to new cases). This training and test data separation was used: (1) to determine the structure of the underlying dimensions of behavior present during test situations, and (2) to determine the relative predictive validity of BR's versus SR's with respect to training certification.

#### 2.5.1. Data analysis aim 1: determination of the underlying dimensions of behavior using both BR and SR methods

Principal components analysis (PCA) and Procrustes analysis were used to address the first aim of our study. We first performed two separate PCA analyses using the two ratings methods, in both cases only on the training data set (training dataset *N* = 251). We tested the sampling adequacy of both PCAs using the Bartlett sphericity test and the Kaiser–Meyer–Olkin (KMO) measure (Budaev, 2010). For the PCA using BR's, we first standardized the single ratings since ratings were given either on a 1–3 scale (object), a 1–4 scale (leash BR and fear BR from the gunfire sub-test), or a 1–5 scale (all other ratings). The correlation matrix of the standardized BR's was deemed appropriate for PCA

(Bartlett's sphericity  $\chi^2_{(300)} = 3228.7$ ,  $P < 0.0001$ ; KMO measure = 0.861). For the PCA using SR's, the correlation matrix was also deemed appropriate (Bartlett's sphericity  $\chi^2_{(78)} = 1419.4$ ,  $P < 0.0001$ ; KMO measure = 0.833). The number of components extracted for each solution matrix was based on a scree test, a parallel analysis, evaluations of simple structure, and the interpretability of the components themselves (Cattell, 1966; Zwick and Velicer, 1986). For component interpretation, variables with a loading of at least  $\pm 0.40$  (rounded to the nearest one-hundredth) were considered to contribute to the meaning of a component (Tabachnick and Fidell, 1996). For both PCAs, orthogonal varimax and oblique direct-oblimin rotated solution matrices were examined, and both methods resulted in an exact (SR data) or similar (BR data) pattern of loadings of measured variables; oblique methods in the case of BR's resulted in the best simple structure, so we report oblique results here for both measurement method types.

We tested the extent to which the factor structure generalized in an independent sample (i.e., the test data set,  $N = 132$ ) for the two types of ratings methods using targeted Procrustes rotation (McCrae et al., 1996). Targeted rotation of solution matrices assesses how the pattern of component loadings obtained for ratings in the training data set were replicated in the test data one. Congruence coefficients greater than 0.85 were taken as evidence of replication across the two matrices (Lorenzo-Seva and ten Barge, 2006). We used SPSS 19.0.0 to implement all PCA and Procrustes procedures.

Aggregate behavior scores were generated based on the pattern of loadings that were obtained in PCA solution matrices. We calculated average unit-weighted scores by averaging the standardized ratings that loaded highly for a particular component identified in PCA. For example, six SR's that clustered with themselves but not with other variables were 'competitiveness', 'hunting drive', 'liveliness', 'cooperation', 'prey drive', and 'curiosity'. Thus, in order to create an aggregate score for the broader dimension identified in PCA, we averaged the ratings given for these six SR's. We computed separate component scores for each dog for each component for each rating method, resulting in eight unique scores per dog (three component scores for SR method, five component scores for BR method). In order for all three SR component scores to have the same meaning of direction, we reverse coded the loadings obtained in PCA3 previous to average weighting the scores. Similarly, in order for all five BR component scores to have the same directionality, we reverse coded the loadings obtained in PCA3 and PCA5 previous to creating aggregate scores. Average unit-weighting was used instead of regression methods to generate scores because PCA loadings are sensitive to sample sizes and to facilitate future attempts at independent study validation (Goldberg and Digman, 1994; Gorsuch, 1983).

## 2.5.2. Data analysis aim 2: prediction of acceptance for training and training completion

We used generalized linear mixed models to predict binary responses from training outcomes (training successfully completed or dog was rejected at some time

during training). For each model the predictor variables included an individuals' aggregate BR and SR scores, its age, the location of the initial test (5 levels in BR models; 4 levels in SR models), and an individual's litter identity fit as a random factor. Two generalized models were fit to the training data, and model parameters from these models were then used to predict outcomes from the independent test data cases to generate a prediction error:

- (1) Model 1: training outcome = BR aggregate score 1 + BR aggregate score 2 + BR aggregate score 3 + BR aggregate score 4 + BR aggregate score 5 + age + sex + location of initial test + litter identity.
- (2) Model 2: training outcome = SR aggregate score 1 + SR aggregate score 2 + SR aggregate score 3 + age + sex + location of initial behavior test + litter identity.

After removal of any dogs that developed medical conditions during training and dogs where training outcomes were pending or unknown, models 1 and 2 training data  $N$  was 118, and test  $N$  was 60. All generalized mixed models were solved by restricted maximum likelihood and implemented in R (Version 2.13.2).

We fit both models first with quadratic terms for all behavioral aggregate scores; in all quadratic models log-likelihood and AIC values were higher compared to identical linear effects only models, so we report linear effects only model results. All models were checked for residual normality and multicollinearity; statistical assumptions were satisfied in all cases.

Overall model significance in each case was based on a log-likelihood ratio test (i.e., compared the fully parameterized model with a reduced model with no behavioral score parameters), and significance of predictors within each model was based on a Wald's test.

We used prediction error (PE) to assess the relative strength of prediction for the two sets of behavioral measurements (BR versus SR). PE for each model was assessed by the percentage of cases in the independent test data that were correctly classified using the model parameters generated from the analysis of training data.

## 2.5.3. Data analysis aim 3: inter-relationships amongst aggregate PCA scores

We used Pearson correlations to estimate the relationships between the five component scores from the BR method and the three component scores from the SR method (all  $N$ 's = 383). Due to the large number of comparisons (Pike, 2011) and large sample size (Yoccoz, 1991) for this analysis, we emphasized interpretation of correlation coefficients as weak/not different from zero (0.00 to  $< 0.30$ ), moderate (0.30–0.50), and strong ( $> 0.50$ ) but report  $P$  values. We also provide a correlation matrix at the level of single BR and SR ratings themselves in the [online supplementary material \(Appendix A\)](#).

**Table 3**

Component loadings of behavioral rating single items on five obliquely rotated principal components. Boldface indicates the highest component loading(s) for each rating. A: acoustic startle sub-test; V: visual startle sub-test; G: gradual visual startle sub-test; GF: gunfire sub-test.

| Principal components:<br>subjective ratings | Confidence  | Physical<br>engagement | Social<br>engagement | Aggression  | Environmental<br>sureness |
|---------------------------------------------|-------------|------------------------|----------------------|-------------|---------------------------|
| Affability                                  | 0.14        | −0.04                  | − <b>0.76</b>        | −0.16       | −0.16                     |
| Handling                                    | 0.12        | 0.14                   | − <b>0.80</b>        | −0.08       | 0.01                      |
| Leash                                       | 0.07        | <b>0.37</b>            | − <b>0.64</b>        | 0.00        | 0.03                      |
| Tug-of-war                                  | −0.07       | <b>0.70</b>            | −0.25                | 0.09        | −0.02                     |
| Chasing                                     | −0.13       | <b>0.78</b>            | −0.20                | 0.01        | −0.11                     |
| Interest in object                          | −0.08       | <b>0.74</b>            | −0.20                | −0.03       | −0.10                     |
| Reaction in dark room                       | 0.21        | 0.02                   | −0.07                | 0.05        | − <b>0.56</b>             |
| Metal stair                                 | 0.09        | 0.12                   | −0.04                | −0.09       | − <b>0.62</b>             |
| Reaction on table                           | 0.04        | 0.14                   | −0.02                | −0.05       | − <b>0.71</b>             |
| Object                                      | −0.02       | <b>0.66</b>            | 0.11                 | −0.12       | −0.31                     |
| A. Flight distance                          | <b>0.38</b> | 0.06                   | 0.03                 | 0.13        | − <b>0.45</b>             |
| A. Secondary response                       | −0.12       | −0.03                  | 0.01                 | 0.06        | − <b>0.46</b>             |
| A. Lasting effect                           | <b>0.52</b> | −0.05                  | −0.03                | 0.11        | − <b>0.41</b>             |
| V. Flight distance                          | <b>0.79</b> | −0.14                  | −0.06                | −0.02       | −0.07                     |
| V. Aggression                               | −0.11       | 0.04                   | 0.08                 | <b>0.86</b> | −0.08                     |
| V. Secondary response                       | <b>0.81</b> | 0.03                   | −0.16                | −0.08       | 0.12                      |
| V. Lasting effect                           | <b>0.87</b> | −0.06                  | −0.04                | −0.06       | −0.05                     |
| G. Fearfulness                              | <b>0.63</b> | 0.10                   | −0.04                | <b>0.37</b> | 0.01                      |
| G. Aggression                               | 0.25        | 0.06                   | 0.07                 | <b>0.74</b> | −0.05                     |
| G. Secondary response                       | <b>0.71</b> | 0.18                   | −0.02                | −0.01       | 0.09                      |
| G. Lasting effect                           | <b>0.78</b> | 0.10                   | 0.05                 | −0.00       | −0.06                     |
| Intensity                                   | 0.23        | <b>0.82</b>            | 0.15                 | 0.06        | 0.10                      |
| Persistence                                 | 0.27        | <b>0.79</b>            | 0.18                 | 0.06        | 0.08                      |
| GF. Fear                                    | <b>0.41</b> | 0.10                   | 0.04                 | 0.10        | −0.11                     |
| Curiosity                                   | −0.06       | −0.07                  | − <b>0.37</b>        | 0.30        | 0.06                      |
| % variance explained                        | 31.2        | 10.4                   | 7.2                  | 5.7         | 5.4                       |

### 3. Results

#### 3.1. Aim 1: underlying dimensions of behavior using BR and SR methods

Five components from the PCA performed on the BRs in the training data were chosen as the best fit of the data, accounting for 60.0% of the variance in the original BR's themselves (Table 3). The component names were chosen based on the definitions of the items that loaded strongly on each component and, where possible, in conjunction with previous research (reviewed in Jones and Gosling, 2005). The first component, which we named 'confidence', consisted of ratings related to reactivity/sensitivity (acoustic startle flight distance, visual startle flight distance, gun fire fear), environmental stability (acoustic startle lasting effect, visual startle lasting effect, gradual startle lasting effect) and investigation (visual startle secondary response, gradual startle secondary response, gradual startle fearfulness). All had high positive loadings on this component. The second component, named 'physical engagement' here, described dogs that varied with respect to their willingness to be engaged in activities related to inanimate objects. The BR's leash, tug-of-war, chasing, interest in objects, object ratings from the unstable table sub-test, search intensity and persistence all loaded positively with one another on the second component. Similarly, component three described a continuum of dogs that varied with respect to their levels of 'social engagement': affability, handling, leash, and curiosity from gunfire sub-tests all covaried with one another. The fourth component, named 'aggression', described differences amongst dogs in their

aggression ratings from visual startle sub-tests, and fearfulness and aggression ratings from gradual visual startle sub-tests. Finally, the fifth component described a continuum of dogs that varied in their 'environmental sureness': ratings from the dark room sub-test, the metal stair sub-test, and the three ratings taken from the acoustic startle sub-test (flight distance, secondary response, and lasting effect) all loaded highly on this final component. Component correlations amongst the five BR components were small to moderate; three correlations neared or exceeded 0.30 (0.33 between PCA1 and PCA2, −0.27 between PCA1 and PCA5, −0.30 between PCA2 and PCA5).

**Table 4**

Component loadings of subjective rating single items on three obliquely rotated principal components. Boldface indicates the highest component loading(s) for each rating.

| Principal components:<br>subjective ratings | Engagement  | Confidence  | Aggression    |
|---------------------------------------------|-------------|-------------|---------------|
| Affability                                  | 0.16        | <b>0.53</b> | <b>0.56</b>   |
| Competitiveness                             | <b>0.81</b> | 0.02        | −0.09         |
| Hunting drive                               | <b>0.88</b> | −0.13       | −0.08         |
| Environmental sureness                      | 0.04        | <b>0.56</b> | −0.20         |
| Courage                                     | 0.11        | <b>0.66</b> | −0.18         |
| Nerve stability                             | −0.09       | <b>0.88</b> | 0.11          |
| Hardness                                    | 0.01        | <b>0.79</b> | −0.08         |
| Liveliness                                  | <b>0.82</b> | −0.05       | −0.02         |
| Sharpness                                   | 0.04        | 0.14        | − <b>0.86</b> |
| Defense drive                               | 0.11        | 0.30        | − <b>0.73</b> |
| Cooperation                                 | <b>0.75</b> | 0.01        | 0.18          |
| Prey drive                                  | <b>0.81</b> | 0.03        | −0.03         |
| Curiosity                                   | <b>0.72</b> | 0.11        | 0.02          |
| % variance explained                        | 37.5        | 15.9        | 11.1          |

**Table 5**

Targeted Procrustes rotation comparing single item behavioral ratings component loadings obtained from training data to those obtained from independent test data. Boldfaced loadings indicate the item the factor loaded on in the training sample.

| Behavioral rating     | Confidence  | Physical engagement | Social engagement | Aggression  | Environ-mental sureness | Item congruence |
|-----------------------|-------------|---------------------|-------------------|-------------|-------------------------|-----------------|
| Affability            | 0.19        | −0.02               | <b>−0.72</b>      | −0.11       | 0.26                    | 0.89            |
| Handling              | 0.18        | 0.15                | <b>−0.78</b>      | −0.03       | 0.11                    | 0.96            |
| Leash                 | 0.10        | <b>0.37</b>         | <b>−0.62</b>      | 0.06        | 0.06                    | 0.99            |
| Tug-of-war            | −0.10       | <b>0.69</b>         | −0.24             | 0.15        | 0.06                    | 0.97            |
| Chasing               | −0.18       | <b>0.77</b>         | −0.20             | 0.07        | 0.14                    | 0.97            |
| Interest in object    | −0.12       | <b>0.73</b>         | −0.19             | 0.03        | 0.13                    | 0.96            |
| Reaction in dark room | 0.18        | 0.02                | 0.03              | 0.05        | <b>0.59</b>             | 0.90            |
| Metal stair           | 0.06        | 0.13                | 0.05              | −0.08       | <b>0.63</b>             | 0.91            |
| Reaction on table     | 0.00        | 0.14                | 0.07              | −0.04       | <b>0.71</b>             | 0.99            |
| Object                | −0.10       | <b>0.66</b>         | 0.14              | −0.09       | 0.30                    | 0.90            |
| A. Flight distance    | <b>0.36</b> | 0.07                | 0.14              | 0.13        | <b>0.46</b>             | 0.89            |
| A. Secondary response | −0.14       | −0.05               | 0.06              | 0.07        | <b>0.44</b>             | 0.65            |
| A. Lasting effect     | <b>0.50</b> | −0.03               | 0.08              | 0.10        | <b>0.43</b>             | 0.77            |
| V. Flight distance    | <b>0.79</b> | −0.08               | 0.02              | −0.05       | 0.10                    | 0.97            |
| V. Aggression         | −0.10       | −0.02               | 0.14              | <b>0.86</b> | 0.06                    | 0.98            |
| V. Secondary response | <b>0.82</b> | 0.10                | −0.10             | −0.09       | −0.06                   | 0.93            |
| V. Lasting effect     | <b>0.87</b> | 0.01                | 0.04              | −0.09       | 0.08                    | 0.82            |
| G. Fearfulness        | <b>0.64</b> | 0.13                | 0.04              | <b>0.36</b> | 0.01                    | 0.93            |
| G. Aggression         | 0.25        | 0.04                | 0.14              | <b>0.74</b> | 0.04                    | 0.97            |
| G. Secondary response | <b>0.70</b> | 0.23                | 0.03              | −0.02       | −0.06                   | 0.84            |
| G. Lasting effect     | <b>0.76</b> | 0.16                | 0.13              | −0.02       | 0.08                    | 0.97            |
| Intensity             | 0.16        | <b>0.84</b>         | 0.16              | 0.09        | −0.10                   | 0.94            |
| Persistence           | 0.20        | <b>0.80</b>         | 0.20              | 0.09        | −0.08                   | 0.98            |
| GF. Fear              | <b>0.39</b> | 0.12                | 0.10              | 0.10        | 0.12                    | 0.59            |
| Curiosity             | −0.01       | −0.10               | <b>−0.36</b>      | 0.32        | −0.01                   | 0.70            |
| Factor congruence     | 0.97        | 0.91                | 0.79              | 0.90        | 0.93                    | 0.91            |

Three components from the PCA performed on the SRs in the training data were chosen as the best fit of the data, accounting for 64.4% of the variance in the original SRs themselves (Table 4). The first component, which we named ‘engagement’, consisted of ratings related to a dog’s willingness to engage with physical and social factors in its environment. The traits competitiveness, hunting drive, liveliness, cooperation and curiosity all loaded high in this factor. The second component, named ‘confidence’ here, described dogs that varied with respect to their affability, environmental sureness, courage, nerve stability and hardness. Similarly, component three described a continuum of dogs that varied with respect to their levels of aggression: sharpness, defense drive and affability (negative loading). There was only one component correlation that exceeded 0.30; engagement and confidence components were moderately and positively correlated with one another ( $r = 0.37$ ).

In the targeted Procrustes rotation of the solution matrix from BR training data to the solution matrix from the test data the overall factor congruence was high (0.91) and the individual component congruences were similarly high for confidence (0.97), physical engagement (0.91), social engagement (0.79), aggression (0.90) and environmental sureness (0.93), indicating a high degree of convergence for the five components across the two independent data sets (Table 5).

In the targeted Procrustes rotation of the solution matrix from SR training data to the solution matrix from the test data the overall factor congruence was excellent (0.97) and the individual component congruences were similarly high for engagement (0.98), confidence (0.97), and aggression (0.95), indicating a high degree of convergence for all three

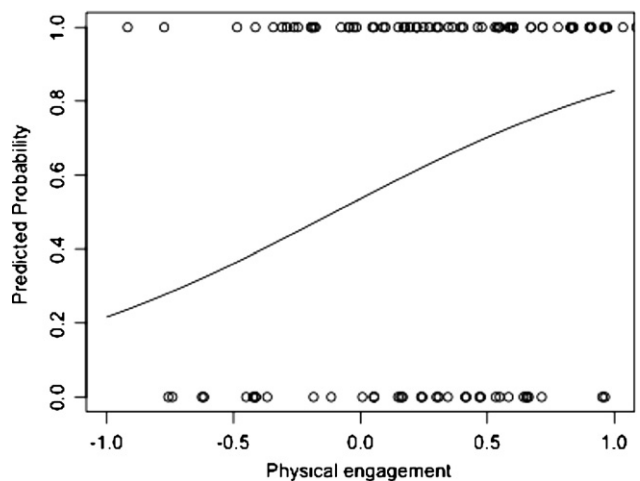


**Fig. 1.** Relationship between physical engagement (an aggregate score based on principal components analysis of the behavioral ratings) and the predicted probability of success in the Swedish armed forces training program. Logistic regression predicted values were generated by holding all other variables in the model (model 1) at their average.

components in the SR data. Similar to the Procrustes results using BR's, SR's tended to load on the same components across training and test data sets, and single SR's that loaded on a particular component in the training data set tended not to load on a different component in the test data set (Table 6).

**Table 6**

Targeted Procrustes rotation comparing single item subjective ratings component loadings obtained from training data to those obtained from independent test data. Boldfaced loadings indicate the item the factor loaded on in the training sample.

| Subjective rating      | Engagement  | Confidence  | Aggression   | Item congruence |
|------------------------|-------------|-------------|--------------|-----------------|
| Affability             | 0.14        | <b>0.52</b> | <b>0.57</b>  | 0.98            |
| Competitiveness        | <b>0.81</b> | 0.04        | −0.08        | 0.92            |
| Hunting drive          | <b>0.88</b> | −0.10       | −0.07        | 0.98            |
| Environmental sureness | 0.03        | <b>0.57</b> | −0.19        | 0.87            |
| Courage                | 0.09        | <b>0.67</b> | −0.17        | 0.96            |
| Nerve stability        | −0.12       | <b>0.88</b> | 0.13         | 0.99            |
| Hardness               | −0.01       | <b>0.79</b> | −0.06        | 1.00            |
| Liveliness             | <b>0.82</b> | −0.03       | −0.01        | 0.98            |
| Sharpness              | 0.05        | 0.16        | <b>−0.86</b> | 0.99            |
| Defense drive          | 0.11        | 0.32        | <b>−0.72</b> | 0.98            |
| Cooperation            | <b>0.74</b> | 0.02        | 0.19         | 0.98            |
| Prey drive             | <b>0.81</b> | 0.06        | −0.02        | 0.99            |
| Curiosity              | <b>0.71</b> | 0.14        | 0.04         | 0.97            |
| Factor congruence      | 0.98        | 0.97        | 0.95         | 0.97            |

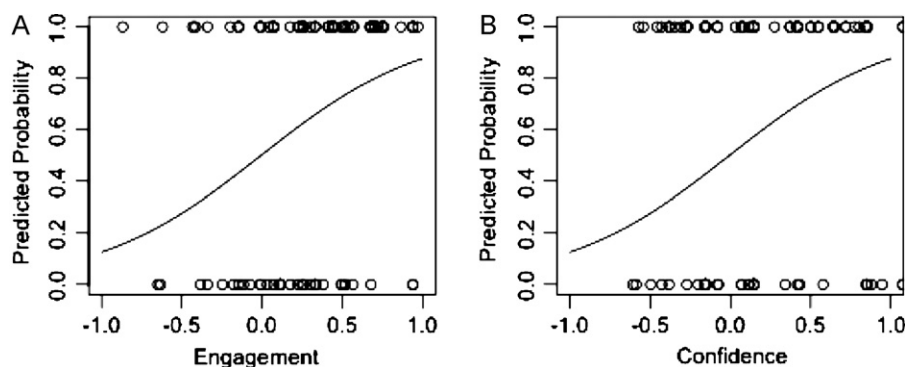


**Fig. 2.** Relationships between engagement (panel A) and confidence (panel B) (aggregate scores based on principal components analysis of the subjective ratings) and the predicted probability of success in the Swedish armed forces training program. Logistic regression predicted values were generated by holding all other variables in the model (model 2) at their average.

### 3.2. Aim 2: prediction of SAF training completion

Behavioral ratings successfully predicted final training outcomes better than a model that included sex, test location, and age only (model 1:  $\chi^2_{(5)} = 21.1$ ,  $P = 0.001$ ; 7.0% variation explained by the random factor litter). However, physical engagement was the only significant BR predictor of an individual dog's final training outcome ( $z = 2.69$ ,  $P < 0.001$ ; Fig. 1). Specifically, the odds of successfully completing SAF training increased 4.2 times with a 1 unit increase in physical engagement (L95%CI = 1.5; U95%CI = 11.8). Using BRs, model 1 correctly classified 72.0% of cases in the training data set, and 78.3% of cases in the independent test data set.

Subjective ratings successfully predicted final training outcomes better than a model that included sex, test location, and age alone (model 2:  $\chi^2_{(3)} = 21.80$ ,  $P < 0.001$ ; 3.9% variation explained by the random factor litter). Engagement ( $z = 3.35$ ,  $P < 0.001$ ) and confidence ( $z = 2.51$ ,  $P < 0.05$ ) were the only significant SR predictors (Fig. 2). Specifically, the odds of successfully completing training in the SAF program increased 7.03 times with a 1 unit increase in engagement (L95%CI = 2.25; U95%CI = 22.02), and 3.43 times with a 1 unit increase in confidence (L95%CI = 1.31; U95%CI = 8.98). Using SRs, model 2 correctly classified 70.3% of final outcomes in the training data set, and

71.7% of final outcomes in the independent test data set.

### 3.3. Aim 3: inter-relationships amongst aggregate PCA scores

There were several moderate to strong relationships across BR and SR component scores (Table 7). Specifically, the SR engagement was moderately related to the BR confidence ( $r = 0.37$ ), moderately related to the BR social engagement ( $r = 0.44$ ), and strongly related to the BR physical engagement ( $r = 0.81$ ). The SR confidence

**Table 7**

Correlations between the five behavioral rating component scores and the three subjective rating component scores in SAF German shepherds. Behavior ratings are given in rows, subjective ratings as columns.  $N = 383$  for all comparisons;  $P$  values given in parentheses.

|                           | SR engagement | SR confidence | SR aggression  |
|---------------------------|---------------|---------------|----------------|
| BR confidence             | 0.37 (<0.001) | 0.70 (<0.001) | 0.26 (<0.001)  |
| BR physical engagement    | 0.81 (<0.001) | 0.38 (<0.001) | 0.03 (0.49)    |
| BR social engagement      | 0.44 (<0.001) | 0.32 (<0.001) | −0.27 (<0.001) |
| BR aggression             | 0.28 (<0.001) | 0.39 (<0.001) | 0.66 (<0.001)  |
| BR environmental sureness | 0.35 (<0.001) | 0.55 (<0.001) | 0.17 (<0.001)  |

was moderately related to the BR's physical engagement ( $r=0.38$ ), aggression ( $r=0.39$ ), and environmental sureness ( $r=0.55$ ), and strongly related to the BR confidence ( $r=0.70$ ). The SR aggression was strongly related to the BR aggression ( $r=0.66$ ), and unrelated to any other BR component (all  $r$ 's  $<0.30$ ).

#### 4. Discussion

In general, there are two basic methods for measuring behavior, behavior codings and behavior ratings (Martin and Bateson, 1993; Vazire et al., 2007). Of the latter, two additional methods can be distinguished, behavior and subjective ratings. The fundamental difference between these two ratings methods is the level of generalizability, or aggregation required by observers, that occurs. In theory, increased generalizability or aggregation of a measure should improve its predictive validity (Epstein, 1983, 1984; Gosling et al., 2003), however, this is also dependent on the domain breadth of the outcome measurement (i.e., a more narrow specific predictor may predict a narrow specific outcome if they both measure the same breadth or construct of interest). Both BR's (e.g., Svartberg et al., 2005) and SR's (e.g., van den Berg et al., 2006) have been used previously in studies of predicting behavior in dogs, but the relative predictive validity of the two ratings methods has not been previously compared within the same sample of dogs, which allows for a direct comparison of the predictive validity of the two methods (for an example from the primate literature, see Uher and Asendorpf, 2008). Here we performed such a direct test. Overall, we found that the two methods appeared to have roughly equal predictive validity with regards to training success in the SAF program. However, there were some fine-scale differences between the two methods, illustrated through several moderate to strong relationships between the 'personality dimensions' identified using each method. Below, we discuss these findings in greater detail and their implications for different measurement methods in this and other working dog programs.

##### 4.1. Relative predictive validity of SR and BR measurement methods

Measurement theory predicts that behavioral measurements that are more broad, or generalizable, should have higher predictive validity due to the fact that observer aggregation will reduce error variance (Epstein, 1983, 1984; Vazire et al., 2007). However, we found that both measurement methods had high, but similar, levels of predictive validity with regards to success in training outcomes (70.3–78.3% correct classifications). Using both methods, the dimensions of confidence and engagement (physical, rather than social) were the strongest predictors of whether a dog passed or failed in the SAF training program. Aggression (both BR and SR methods), the BR environmental sureness, and the BR social engagement in both cases had little, if any, predictive validity. Taken at face value, our results suggest that in some cases, from the standpoint of predictive validity, the choice of measurement method (BR or SR) may have little consequence.

However, as with any scientific study, alternative explanations exist. For example, how different were the two measurement methods that were used here? In other words, our subjective ratings may have been restricted in their breadth, since they were based on only one test occasion and the test leader had no previous knowledge of the dogs (see also Gosling et al., 2003). Correlations between SR and BR components indicated that each SR component was at least strongly related to one BR component, and with the exception of the SR aggression, moderately related to other BR components. Broader subjective ratings, aggregated across multiple tests and observations could improve the predictive validity of an SR method relative to more restricted BR's based on specific sub-test observations. In essence, due to the logistical constraints of the SAF program, the test leader's domain of knowledge may not have been generalizable enough to fully enable SR methods to capture greater amounts of domain breadth of the outcome variable (success in training).

Another potential alternative explanation is that the breadth of the outcome domain was narrow. In this case, use of a more narrow, specific predictor (e.g., a BR) could result in equality of predictive validity with a more broad, generalizable one (e.g., a SR) if there is a close match between the domain of behavior measured by the narrow predictor and the domain of behavior encapsulated by a narrowly defined outcome variable. Indeed, while many working dog programs utilize standardized behavioral tests to quantify continuous behavioral variation (i.e., attempt to measure a broad domain of overall behavior), these same programs tend to have dichotomous outcome variables as their success criteria (i.e., success/failure), which may fail to quantitatively capture precisely what exact aspects of behavior result in a success or failure outcome. Interestingly, to the best of our knowledge, characterizing domains of success outside of simple dichotomous outcomes has not been performed in most working dog programs. This may be a fruitful area of future research, and may also allow for more broad, generalizable predictors (e.g., SR methods) to have stronger predictive characteristics. Further work on characterizing the domains of behavior that are captured during both standardized testing as well as in the determination of success criteria is needed to test these ideas.

The BR physical engagement classified 6.6% of dogs more accurately than the SRs confidence and engagement with regards to training outcome. In this case, this result implies that, in the SAF program: (1) access to SR ratings (but not BR ones) by SAF training staff appears to not overly influence decisions to reject/certify a dog from training, (2) training outcomes may be a result of narrowly defined ranges of appropriate behavior, and (3) that these more narrow ranges of behavior can be successfully captured by the BR physical engagement alone.

##### 4.2. Inter-relationships amongst aggregate scores generated by BR and SR methods

The first SR component and the second and third BR components, named engagement, physical engagement, and social engagement, respectively, were either strongly

(the SR engagement and the BR physical engagement) or moderately (the SR engagement and the BR social engagement) correlated with one another across the sample of dogs tested here. The BRs that contributed to an individual's physical engagement included initiating and maintaining control of inanimate objects, while the BRs that contributed to an individual's social engagement included ratings associated with responsiveness and attention to a handler and other people. The SR engagement tended to capture both of these types of engagement (as well as aspects of an individual's curiosity and energy levels), illustrating how use of more general, broad ratings (SR's) may sometimes result in a loss of more specific information with regards to any specific situation. However, in this case the broader aggregation in SR engagement did not result in meaningful differences in predictive validity between the two measurement methods. In the case of the SAF program then, choice of measurement method for human and physical engagement may be irrelevant (i.e., SR methods or BR methods both seem to capture relevant aspects of engagement equally well) since positive, high scores for social and physical engagement are considered desirable.

This may not be the case, however, for all working dog programs. On the one hand, similar levels of engagement with objects and humans should be important for working dogs whose tasks include explosives or narcotics detection (Maejima et al., 2007; Sinn et al., 2010; Svartberg, 2002), so using a general, broad SR method that does not distinguish types of engagement may be the least time consuming and most reliable method of measurement (Vazire et al., 2007). On the other hand, in some working dogs, social engagement may be considered desirable while physical engagement may not (i.e., a negative relationship is desirable, for example, in guide dogs; Goddard and Beilharz, 1986). In these cases, working dog programs may need to consider how inherent levels of aggregation of behavior in their measurement method may influence their ability to choose candidates based on appropriate domains of behavior. If finer scale differences are needed (e.g., one aspect of engagement is desired, but another is not), programs may need to choose to use BR methods at the expense of measurement breadth.

The second SR component and the first BR component were also strongly correlated and in both cases captured aspects of a dog's 'confidence'. Confidence and other aspects of sensitivity to disturbances have also been identified as a broad factor describing important individual differences in companion dogs ('fearlessness': Svartberg and Forkman, 2002), police and detection dogs ('mental stability': Wilsson and Sundgren, 1997a), and guide dogs ('fearfulness': Goddard and Beilharz, 1986; Pfaffenberger et al., 1976). The SR confidence also encapsulated aspects of an individual's physical engagement, aggression, and environmental sureness (all BR components that had moderate (~0.40) correlations with the SR confidence). Perhaps not surprisingly, aggression in dogs can be closely related to fearfulness, or lack of confidence (Bräm et al., 2008; Haverbeke et al., 2009), and environmental sureness as defined in the BR component included what could be considered confidence in strange environments. The SR confidence was an important predictor of training success,

but the BR confidence was not; in this case the measurement breadth of a trait induced by measurement methodology did appear to result in differences in predictive validity. Again, the point here is that working dog programs may need to carefully consider the trade-off between the use of SR methods, where aggregation and broad generalization should improve predictive validity, and the use of BR methods, which may be able to identify more situation-specific traits that are of utmost necessity for a particular type of work or companionship.

The third SR component and the fourth BR component, both named aggression here (Svartberg and Forkman, 2002; van der Borg et al., 2010; Wilsson and Sundgren, 1997a), were strongly correlated with one another, and the SR aggression was not related to any other BR factors. Aggression has also been found to be largely independent of other personality traits in other studies (Svartberg, 2002; Svartberg et al., 2005). From the standpoint of measurement method, it appeared that either BR or SR techniques are equally likely to distinguish between individual dogs in their levels of aggressiveness.

With regards to our prediction models (our second aim), independent predictors (BR and SR scores) were measured by a different individual than those responsible for measuring the dependent outcome (success or failure in training). However, it is worth noting that this was not the case with regards to our results on the inter-relationships of the scores generated by the two methods, in which all scores were generated by the same person. Therefore, our inference above regarding inter-relationships of dimensions of behavior measured by SR and BR methods is limited in that there is the possibility that our results are relevant to this specific test leader in the SAF program only. It is also worth noting that, while not guaranteed, reports indicate that inter-rater reliability estimates for most measured behaviors are acceptable/high in most dog studies (Rooney et al., 2007; Sinn et al., 2010) as well as in other animals (Gosling and Vazire, 2002). Still, establishing both inter- and intra-rater reliability (Taylor and Mills, 2006) would broaden the level of inference of any further research in the SAF program. Replication of this result in other working dog programs awaits validation.

## 5. Conclusions

Despite the implications of measurement theory (John and Soto, 2007), our study suggests that from the standpoint of predictive validity, use of different ratings measurement methods (BR versus SR) sometimes may not matter, and may largely be a matter of choice (or logistics) by researchers or working dog programs. However, narrower, more situation-specific traits (e.g., specific aspects of engagement) may be necessary to capture for a particular working dog program in some cases, while for other types of programs or traits (such as general confidence or aggression) this may not be the case, and more generalizable measures (i.e., SRs) are more efficient and logistically feasible. This emphasizes the need for more explicit, quantitatively continuous definitions of 'success' in working dog programs – without knowing the breadth and domain of specific traits that are important for success, it may be

difficult for working dog programs to make choices regarding the measurement methods that they use to capture relevant behavioral variation observed in their dogs earlier in life.

## Acknowledgements

Sam Gosling provided advice on nomenclature and factor analytic structure of tests. Nate Marti provided statistical assistance. Pierre Wahlström performed all of the actual testing and was responsible for test design. Margareta Löw provided critical assistance and support during the tests themselves. Two anonymous reviewers provided helpful feedback on a previous version of this manuscript.

## 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.applanim.2012.08.012>.

## References

- Block, J., 1961. The Q-sort method in personality assessment and psychiatric research. Charles C. Thomas, Springfield, IL.
- Bräm, M., Doherr, M.G., Lehmann, D., Mills, D., Steiger, A., 2008. Evaluating aggressive behavior in dogs: a comparison of 3 tests. *J. Vet. Behav.: Clin. Appl. Res.* 3, 152–160.
- Budaev, S.V., 2010. Using principal components and factor analysis in animal behaviour research: caveats and guidelines. *Ethology* 116, 472–480.
- Cattell, R.B., 1966. The scree test for the number of factors. *Social. Method. Res.* 1, 245–276.
- De Meester, R.H., De Bacquer, D., Peremans, K., Vermeire, S., Planta, D.J., Coopman, F., Audenaert, K., 2008. A preliminary study on the use of the socially acceptable behavior test as a test for shyness/confidence in the temperament of dogs. *J. Vet. Behav.: Clin. Appl. Res.* 3, 161–170.
- Diederich, C., Giffroy, J.-M., 2006. Behavioural testing in dogs: a review of methodology in search for standardization. *Appl. Anim. Behav. Sci.* 97, 51–72.
- Epstein, S., 1983. Aggregation and beyond: some basic issues on the prediction of behavior. *J. Pers.* 51, 360–392.
- Epstein, S., 1984. The stability of behavior across time and situations. In: Zucker, R., Arnoff, J., Rabin, A.I. (Eds.), *Personality and the Prediction of Behavior*. Academic Press, San Diego, CA, pp. 209–268.
- Goddard, M.E., Beilharz, R.G., 1986. Early prediction of adult behaviour in potential guide dogs. *Appl. Anim. Behav. Sci.* 15, 247–260.
- Goldberg, L.R., Digman, J.M., 1994. Revealing 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*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Gosling, S.D., 1998. Personality dimensions in spotted hyenas (*Crocuta crocuta*). *J. Comp. Psychol.* 112, 107–118.
- 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, S.Y., John, O.P., 2003. A dog's got personality: a cross-species comparative approach to evaluating personality judgements. *J. Pers. Soc. Psychol.* 85, 1161–1169.
- Gosling, S.D., Vazire, S., 2002. Are we barking up the right tree? Evaluating a comparative approach to personality. *J. Res. Pers.* 36, 607–614.
- 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.
- Hewson, C.J., Luescher, U.A., Ball, R.O., 1998. Measuring change in the behavioral severity of canine compulsive disorder: the construct validity of categories of change derived from two rating scales. *Appl. Anim. Behav. Sci.* 60, 55–68.
- Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *J. Am. Vet. Med. Assoc.* 223, 1293–1300.
- Humphrey, E.S., 1934. "Mental tests" for shepherd dogs. *J. Hered.* 25, 129–136.
- 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.
- Ley, J.M., Bennett, P.C., Coleman, G.J., 2009. A refinement and validation of the Monash canine personality questionnaire (MCPQ). *Appl. Anim. Behav. Sci.* 116, 220–227.
- Lorenzo-Seva, U., ten Barge, 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., 2007. Traits and genotypes may predict the successful training of drug detection dogs. *Appl. Anim. Behav. Sci.* 107, 287–298.
- Martin, P., Bateson, P., 1993. *Measuring Behaviour: An Introductory Guide*, 2nd ed. Cambridge University Press, Cambridge.
- Martínek, Z., Lát, J., Hartl, K., 1971. About the possibility of predicting performance of adult guard dogs from early behaviour. *Activ. Nerv. Super.* 13, 149–150.
- McCrae, R.R., Zonderman, A.B., Costa, P.T., Bond, M.H., Paunonen, S.V., 1996. Evaluating replicability of factors in the revised NEO personality inventory: confirmatory factor analysis versus Procrustes rotation. *J. Pers. Soc. Psychol.* 70, 552–566.
- Netto, W.J., Planta, D.J.U., 1997. Behavioral testing for aggression in the domestic dog. *Appl. Anim. Behav. Sci.* 52, 243–263.
- Pike, N., 2011. Using false discovery rates for multiple comparisons in ecology and evolution. *Methods Ecol. Evol.* 2, 278–282.
- Pfaffenberger, J.C., Scott, J.P., Fuller, J.L., Ginsburg, B.E., Bielefeld, A.W., 1976. *Guide Dogs for the Blind: Their Selection Development and Training*. Elsevier, Amsterdam, The Netherlands.
- 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.
- Ruefenacht, S., Gebhard-Heinrich, 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.
- Saetre, P., Strandberg, E., Sundgren, P.-E., Pettersson, U., Jazin, E., Bergström, T.F., 2006. The genetic contribution to canine personality. *Genes Brain Behav.* 5, 240–248.
- Sih, A., Bell, A.M., Johnson, J.C., Ziemba, R.E., 2004. Behavioral syndromes: an integrative overview. *Q. Rev. Biol.* 79, 242–277.
- 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.
- 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., Forkman, B., 2002. Personality traits in the domestic dog. *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.
- Svartberg, K., 2007. Individual differences in behaviour – dog personality. In: Jensen, P. (Ed.), *The Behavioural Biology of Dogs*. CAB International, Cambridge, MA, pp. 182–206.
- SWGDOG. Retrieved January 11, 2012 at: <http://www.swgdog.org/>.
- Tabachnick, B.G., Fidell, L.S., 1996. *Using Multivariate Statistics*, 3rd ed. Harper Collins, New York.
- Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. *J. Vet. Behav.: Clin. Appl. Res.* 1, 94–108.
- Uher, J., Asendorpf, J.B., 2008. Personality assessment in the great apes: comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *J. Res. Pers.* 42, 821–838.
- Vazire, S., Gosling, S.D., Dickey, A.S., Schaprio, S.J., 2007. Measuring personality in nonhuman animals. In: Robins, R.W., Fraley, R.C., Krueger, R. (Eds.), *Handbook of Research Methods in Personality Psychology*. New York: Guilford, Guilford, New York, pp. 190–206.

- van den Berg, L., Schilder, M., de Vries, H., Leegwater, P., van Oost, B., 2006. Phenotyping of aggressive behavior in golden retriever dogs with a questionnaire. *Behav. Genet.* 36, 882–902.
- van der Borg, J.A.M., Beerda, B., Ooms, M., De Souza, A.S., van Hagen Kemp, B., 2010. Evaluating of behaviour testing for human directed aggression in dogs. *Appl. Anim. Behav. Sci.* 128, 78–90.
- van der Waaij, E.H., Wilsson, E., Strandberg, E., 2008. Genetic analysis of a Swedish behavior test on German shepherd dogs and Labrador retrievers. *J. Anim. Sci.* 86, 2853–2861.
- Willis, M.B., 1995. Genetic aspects of dog behaviour with particular reference to working ability. In: Serpell, J.A. (Ed.), *The Domestic Dog: Its Evolution, Behaviour and Interactions with People*. Cambridge University Press, New York, pp. 51–64.
- Wilsson, E., Sundgren, P.-E., 1997a. The use of a behavior test for the selection of dogs for service and breeding: I. Methods 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., 1997b. The use of a behaviour test for selection of dogs for service and breeding: II. Heritability for tested parameters and selection of dogs for service and breeding. *Appl. Anim. Behav. Sci.* 54, 235–241.
- Yoccoz, N.G., 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull. Ecol. Soc. Am.* 72, 106–111.
- Zwick, W.R., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. *Psychol. Bull.* 99, 432–442.