Applied Animal Behaviour Science 179 (2016) 82–94

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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

Comparing the predictive validity of behavioral codings and behavioral ratings in a working-dog breeding program

Monica E. McGarrity[a], David L. Sinn[a,b,*], Scott G. Thomas[c,e], C. Nathan Marti[d], Samuel D. Gosling[a,e]

[a] University of Texas at Austin, Department of Psychology, 108 E. Dean Keeton Stop A8000, Austin, TX 78712, USA
[b] University of Tasmania, School of Biological Sciences, Private Bag 5, Hobart, TAS 7001 Australia
[c] Department of Homeland Security, Transportation Security Administration, Canine Breeding and Development Center, Lackland Air Force Base, San Antonio, TX, USA
[d] Abacist Analytics, P.O. Box 11581, Austin, TX 78711, USA
[e] School of Psychological Sciences, University of Melbourne, Parkville, VIC, Australia

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form 12 March 2016
Accepted 17 March 2016
Available online 19 March 2016

Keywords:
Working dog
Personality
Measurement methods
Predictive validity
Behavior rating
Behavior coding

ABSTRACT

Most working-dog breeding programs have a substantial interest in using behavioral assessments of their young dogs to predict their subsequent success. Different methods of measuring behavior may capture different aspects of behavior yet working-dog programs typically use only a single measurement method. Thus, the primary aim of this study was to test whether two different measurement methods (ratings or codings) would differ in their predictive validity with respect to working-dog selection outcomes. Rating methods require observers to intuitively aggregate their observations into a single rating and in doing so may reduce error variance in measurement, resulting in improved validity. Coding methods on the other hand do not demand so much judgment on the part of the observer so may be less influenced by observer biases. Here we analyzed the two methods with respect to their ability to predict selection for training in a sample of odor-detection dogs bred at the U.S. Transportation Security Administration Canine Breeding and Development Center. Behaviors observed in two standardized tests (*search & retrieve* and *environment*) at four different time points across the first year of life were measured using nine ratings and 23 codings. Data reduction techniques identified two underlying dimensions in ratings (environmental stability and hunt drive) and nine in codings (confidence, anxiety, exploration, excitability, search performance, dominant possession, independent possession, energy management, and search aptitude). There were no differences in predictive validity between the two methods; both ratings and codings correctly classified a high percentage of dogs that were/were not selected for training at 12 months of age (84.6–88.5%). In the search & retrieve test, codings and ratings appeared to be measuring the same construct. In the environment test the only significant coding predictor of training selection (confidence) was strongly related to the single rating predictor (environmental stability). Rating methods tended to capture behavior that was more consistent, while coding methods tended to capture behavior that was more situation-specific. Our mixed-models approach also allowed us to discriminate between average behavior (between-individual variation) and behavioral change through time (within-individual variation); such findings emphasize different aspects of development that may need to be monitored during rearing. Our results suggest that, in some cases, the use of ratings versus codings may be inconsequential from the standpoint of predicting which dogs get selected for training. Virtually all research on animal behavior assesses behavior via coding or rating methods; further work is needed to verify these results.

1. Introduction

Detection dogs are used for a wide variety of purposes, including identifying dangerous substances, assisting in conservation actions, and aiding in search and rescue (Helton, 2009). However, individual dogs vary considerably in their general disposition or 'personality', and it is an individual's personality that may play a large part in

* Corresponding author at: University of Texas at Austin, Department of Psychology, 108 E. Dean Keeton Stop A8000, Austin, TX 78712, USA.
E-mail address: psuocto@yahoo.com (D.L. Sinn).

http://dx.doi.org/10.1016/j.applanim.2016.03.013
0168-1591/

determining working success in later life (Goddard and Beilharz, 1982; King et al., 2012; Serpell and Hsu, 2001). Personality can be defined as individual differences in behavior that are correlated across time, functional contexts, or both (Gosling, 2001; Sih et al., 2004; Svartberg, 2007). For example, fearful behaviors in a dog in response to strange humans can be correlated across time, and fearfulness towards strangers may also be correlated with lack of confidence in novel environments. As a result, detector dogs are often selectively bred based for personality traits relevant for their specific working roles. Still, individual differences in personality are often prominent even within these artificially selected populations (Fratkin et al., 2013; Graham and Gosling, 2009; Jones and Gosling 2005).

Research on detector dogs is nascent, but two observations can be made. First, breeding and training programs for working canines are costly. Second, many puppies produced by breeding programs are not 'successful'; normally only 30–50% of all dogs bred end up serving in the roles for which they were bred and raised (Maejima et al., 2007; McGarrity et al., 2012; Slabbert and Odendaal, 1999; Wilsson and Sundgren, 1997). Therefore, efforts to quantify variation in behavior and determine how personality relates to working success are often a high priority. Standardized behavior surveys such as the Canine Behavioral Assessment and Research Questionnaire, or C-BARQ© (http://www.cbarq.org; Duffy and Serpell, 2008, 2012; Hsu and Serpell, 2003) and behavior test batteries (Goddard and Beilharz, 1986; Tomkins et al., 2011; Wilsson and Sundgren, 1997) have been used by many working-dog programs for this purpose (Maejima et al., 2007; Rooney et al., 2004; Sinn et al., 2010; Slabbert and Odendaal, 1999; Wilsson and Sundgren, 1998).

However, most studies use only a single measurement approach to measure dog personality—either a rating or a coding technique (Jones and Gosling, 2005; Vazire et al., 2007). Rating methods require human observers to aggregate their impressions of dog behavior using a Likert-type scale. Rating scores could indicate frequency of a specific behavior (e.g., rarely, sometimes, often) or the degree to which a relatively broad trait is exhibited (e.g., not confident, somewhat confident, extremely confident). In contrast to ratings, which inherently involve aggregation by the observer, coding methods quantify discrete behaviors (e.g., barking) using measures such as frequency counts, duration, or latency (e.g., Batt et al., 2008; Netto and Planta, 1997). Rating and coding measures of the behavior of individual animals can often be strongly correlated (Capitanio, 1999; De Meester et al., 2008; Hewson et al., 1998; Lloyd et al., 2007; van den Berg et al., 2006; Vazire et al., 2007). Indeed, some popular assessments that use ratings, such as the Dog Mentality Assessment (Svartberg, 2002) and the Puppy Behavior Test (Campbell, 1975), have been successfully converted into a coding format (Batt et al., 2008; Beaudet et al., 1994). However, in some cases, scores from the two different measurement methods may not converge (Freeman et al., 2011; Kubinyi et al., 2015; Vazire et al., 2007). Lack of convergence may occur if one or both measures are not reliable (the measures are not reproducible across items, time, or observers) or if one or both measures are not valid (the measures are not tapping the construct or behavior they were designed to measure).

Theoretically, arguments can be made in favor of the superiority of either method in terms of reliability and validity (Vazire et al., 2007). Ratings are often designed in a manner that involves a high degree of inference by the observer (e.g., adjective-based ratings such as "curious") and so they may have a greater chance to capture relevant behavior (i.e., have greater measurement breadth; Uher and Asendorpf, 2008). For example, an observer may derive a single rating score by combining the totality of their experiences of the focal animal's behavior with their previous experiences of behavior in that species or population as a whole. From a psychometric perspective, the aggregation process inherent to ratings also theoretically reduces random error variance in measurements, thereby improving reliability. However, the intuition inherent to ratings may make them more susceptible to systematic rater biases, resulting in less reliable scores (Svartberg and Forkman, 2002). Codings, on the other hand, are often thought to be more objective, thereby reducing the potential effects of rater biases and improving reliability. Codings tend to be more situation-dependent than ratings, and thus capture finer-grained information across shorter periods of time, reducing measurement breadth while also potentially increasing error variance. From a practical standpoint, ratings methods tend to take less time and effort to deploy than codings (Vazire et al., 2007).

Empirical tests comparing the reliability of the two measurement methods indicate that both methods can be reliable (Carter et al., 2012; Fratkin et al., 2013; Gartner and Powell 2012; Gosling 2001; but see Highfill et al., 2010). Nonetheless, even when both methods are reliable they sometimes still fail to converge strongly (Freeman et al., 2011; Kubinyi et al., 2015). In such cases, which method is to be preferred? In most working-dog research and other applied contexts, behavior is measured with the goal of predicting some future outcome. Therefore, one sensible criterion with which to evaluate the two methods is with respect to their relative ability to predict important 'real-life' outcomes (Vazire et al., 2007; Wilsson and Sinn 2012). Unique behavioral variation captured by a particular method could result in different measurement methods being able to differentially predict the capacity of a dog to perform in a given working role.

We addressed the question of predictive-validity differences between ratings and codings by developing coding methods for standardized behavior tests used by the U.S. Department of Homeland Security's Transportation Security Administration Canine Breeding and Development Center's (TSA-CBDC) puppy program; behaviors were already being measured using ratings by the TSA-CBDC during standardized tests. The goal of the TSA-CBDC is to produce a population of dogs with a high overall rate of selection by the TSA Canine Training and Evaluation Section (TSA-CTES); the TSA-CTES are responsible for training odor-detection dogs for real-life work. Use of an appropriate measurement instrument during breeding and development is key to achieving this aim. We evaluated the degree to which behaviors assessed by ratings and coding methods converged and their relative ability to predict dogs' selection for training in the TSA program.

2. Materials and methods

2.1. Subjects

From 2002–2012, the TSA-CBDC bred and reared explosives detection dogs based on a combination of observed behavior, medical requirements, breeding demand, and maintenance of genetic variation. Initial founders of the population were obtained from the Australian Customs and Border Protection Service; subsequent breeding females were produced by the TSA-CBDC or procured from various sources (e.g., US Department of Defense). Stud males were either bred by TSA-CBDC or obtained from privately owned field champions. The majority of dogs produced by the TSA-CBDC were Labrador retrievers but some crossbreeds and Vizslas were also produced. At 8 weeks of age, dogs were typically fostered to members of the public who house-trained the dogs and were encouraged to expose the dog to a variety of environments (e.g., shopping-mall parking areas); however, 'puppy raisers' did not attempt any specialized training beyond seeking to consistently associate physical or verbal praise with toy play. At three, six, nine, and 12 months of age, dogs were returned to the TSA-CBDC for

standardized behavioral testing using a testing protocol developed previously (Champness, 1996, 2000).

The TSA-CTES evaluated dogs at 12 months of age to determine acceptance into detection training. Like all potential criterion measures, this one has its own set of statistical, conceptual, and methodological pros and cons; we discuss these pros and cons in Section 4.3. In line with previous studies, which have used evaluations within their breeding and training programs (e.g., Maejima et al., 2007; Sinn et al., 2010; Wilsson and Sinn, 2012), we use acceptance into training as the criterion we aimed to predict. The final TSA-CTES evaluation consisted of search exercises in a non-novel environment. Decisions to accept or decline a dog for training were made by the TSA-CTES trainers rather than TSA-CBDC staff, although TSA-CBDC staff input was requested in some cases.

Prior to the 12-month final evaluation, dogs could be eliminated by the TSA-CBDC for medical or behavioral reasons. Subject dogs received general weekly veterinary checks, specific checks on a canine development schedule (e.g., hip dysplasia checks at six months) and daily checks by kennel staff.

From 2010-11-03 to 2013-04-15, we collected video of standardized tests for as many dogs as possible (106 dogs) using either a Samsung HD 1030p handheld video camera or a GoPro HERO headcam worn by a tester. We considered two *a priori* conditions for inclusion of dogs in our study. First, our statistical analyses involved estimation of both an 'average' behavior as well as behavioral change or plasticity for each dog so we required at least three data points (i.e., test ages) per dog for inclusion in our study. Second, our validity-criterion variable (i.e., selection for training) required that each dog in our sample receive a final evaluation. For dogs that were videoed at fewer than three test ages (54 dogs), we used their video for pilot testing to develop ethograms for behavioral codings. Two dogs with sufficient video were excluded based on our second *a priori* condition; both dogs exhibited signs of extreme stress during testing on multiple occasions and it was deemed that retaining them in the development program would be unethical. Both dogs were removed and placed in adoptive homes without receiving a final evaluation.

Thus, our sample for analysis consisted of 52 dogs from 13 litters, with an equal number of males and females. The breeds of the 52 dogs included in our sample were as follows: Labrador retriever (N = 40), Vizsla (N = 8), and German shorthaired pointer/Labrador mix (N = 4). Of the 52 dogs in our sample, 83% (43) were accepted for training after the final evaluation; this strongly positive selection outcome is consistent with a previous estimate from a larger sample of TSA-bred dogs (82%; McGarrity et al., 2012). The 'overall success' rate for this population (30% of all dogs bred/procured by the TSA-CBDC ultimately certify for real-life working roles; McGarrity et al., 2012) is similar to overall success rates reported from other working-dog programs (Maejima et al., 2007; Slabbert and Odendaal, 1999).

2.2. Standardized behavior test procedure

The TSA-CBDC standardized behavior tests consisted of an 'environment' (ENV) test and a 'search & retrieve' (S&R) test. Test order was the same for each dog within an age (except for rare cases dictated by weather or staffing), but test order was different across ages. Behavior tests were performed during the day between 09:00 and 17:00 and dogs were usually given both tests on the same day; in the event of inclement weather, testing resumed on the following day. Each test evaluation was typically completed within five to 10 min; however, the complexity and duration of the ENV test were intended to increase with age. The order in which dogs were chosen for testing within any given day was based on transport trailer loading and unloading order, in order to minimize the length of time each dog was confined to its trailer space. An exception to this procedure was that any females immediately pre- or post-estrus were tested last to minimize the distractions to other dogs; dogs in estrus were not tested or were tested at a later date.

The standardized ENV test consisted of a walk through a stimulating environment with the TSA-CBDC test leader. During the walk, the test leader (TL) encouraged the dog to sniff or climb/jump onto a variety of novel objects, presented the dog with challenges such as stairs, slick floors, and automatic doors, and engaged the dog in one or more bouts of toy play. The testing location and associated stimuli varied with age. At three months, the TL walked the dog through a general store with a garden center and busy outdoor sidewalk/entry area. At six months, the TL walked the dog through a busy woodworking shop with loud noises and dark enclosed areas. At nine months, the TL walked the dog through an airport cargo storage area with suitcases, baggage carts, other equipment, cargo-moving traffic, and noise from adjacent runways. At 12 months, the TL walked the dog through the passenger-loading zone, parking areas, and baggage claim area at an international airport.

The S&R test consisted of retrieving and searching exercises in an open field area. With the exception of the 3-month test, the TL attached the dog to a 30-foot leash that allowed the handler to restrain the dog if necessary. Initially, the TL released the dog to run or walk across the field with the TL and a test assistant (TA). Upon reaching the far end of the field, the TL restrained the dog while the TA enticed the dog with a toy (i.e., an explosive-scented rolled towel) and then threw the toy; the TL released the dog to chase and retrieve the toy. Upon retrieval, the TL or TA engaged the dog in a brief game of tug-of-war and then regained possession of the toy. Next, the TL restrained the dog while the TA enticed the dog with the toy and then threw the toy into an area of tall grass; the TL released the dog to search for the toy. When the dog found the toy, the TL and TA immediately began to walk toward the other end of the field, verbally encouraging the dog to follow if necessary but without engaging in toy play. Upon reaching the opposite end of the field, the TL restrained the dog as the TA enticed the dog with the toy and then hid the toy under one of ten upside down flower pots arranged in a line perpendicular to wind direction. The TA remained at the opposite end of the line of pots while the TL released the dog to search for the toy. If the dog persisted in searching—even without finding the toy—it was allowed to continue unassisted; if the dog became distracted from the search, the TL and TA would step into the line of pots, verbally encourage the dog to search, tap the pots to entice investigation, or even move the pot briefly to expose the toy. Next, the hidden pots search was repeated, with the difference that the TA returned to stand beside the TL at the start of the line of pots. At the end of the second search, the dog was encouraged to carry the toy back to the dog trailer but was not engaged in toy play.

2.3. Behavioral measurement procedure

2.3.1. Behavior ratings

All behavioral ratings (BRs) used in this study were part of the already existing TSA-CBDC standardized testing operational protocols. Upon completion of each ENV test, the TL immediately gave ten behavioral ratings (BRs) based on observed overall test performance and toy play. Upon completion of the S&R test, the TL and TA consulted to immediately give eight BRs. Previous analysis of a much larger sample of TSA-CBDC dogs (Sinn et al., 2011) indicated that six of the BRs from the ENV test were invariant, that is, six BRs did not capture variation among individuals. In the same previous analysis of a much larger sample (Sinn et al., 2011), three of the S&R test BRs were invariant, and unable to be used for statistical analysis. The same nine BRs (six from ENV and three from S&R) were invariant in our sample, so we excluded them and used only the nine remaining BRs that captured variation

Table 1
Operational definitions of behavioral ratings given during TSA-CBDC standardized behavioral tests. Each trait is rated from 1 to 5, with 1 representing low expression of a trait and 5 representing high expression of the trait. ENV = environment test; S&R = search and retrieve test.

Test	Trait	Description
ENV	Confidence	An environmentally conditioned acceptance of safety, in other words, a measure of the lack of fear (high fear = low confidence). Typical signs of fear include extended attention, freezing, and avoidance
ENV	Concentration	The dog's focus on the area of search; a measure of the lack of distraction towards objects not related to searches. Typical distraction objects are sticks and leaves on the ground, people, self, and leash biting
ENV	Responsiveness	The dog's ability to react to corrections or encouragement (verbal and physical praise) from the handler
ENV	Initiative	A dog's willingness to walk at the end of the leash and investigate the environment on his/her own without being asked by the handler
S&R	Physical Possession	The dog's desire to play tug-of-war with the handler for the toy, including the force and determination to maintain its grip on the toy
S&R	Mental Possession	Tendency to retain focus on the hiding of the towel. Signs of lower mental possession include the dog focusing on the handler more than the towel
S&R	Independent Possession	The dog's willingness to continue to interact and possess the toy independently of the handler
S&R	Hidden One	The dog's concentration, willingness, and ability to move purposefully down a line of upside down flower pots, one of which contains a hidden scented towel. Higher scores are also given to dogs that show a behavioral change when at the scent cone, and then self-reward with the found scented towel
S&R	Hidden Two	Identical scores are given during a second hunt search at the flowerpots, when the tester stands at a different location previous to the first hunt search (i.e., at the level of the dog/handler, as opposed to at the end of the flower pot line)

between individuals for all subsequent analyses (Table 1). The nine BRs used here have high inter-rater reliability (ENV test: average ICC(3,1) = 0.82; L95%C.I. = 0.53, U95%C.I. = 0.92; S&R test: average ICC(3,1) = 0.89, L95%C.I. = 0.76, U95%C.I. = 0.95; Fratkin et al., 2015).

2.3.2. Behavior codings

In an independent sample (54 dogs), we used pilot videos of TSA-CBDC standardized tests to develop a comprehensive ethogram of 43 behaviors that could be coded from test videos using Scribe 4 (University of Texas at Austin Butler School of Music Center for Music Learning; available at: http://cml.music.utexas.edu/online-resources/scribe-4/description/). We coded behaviors as duration/latency, frequency, binary occurrence (i.e., yes/no), or proportion occurrence. For codings that were recorded as proportion occurrence, we used either segment-occurrence sampling or instantaneous sampling (Martin and Bateson, 1993). To generate a proportion occurrence using segment occurrence sampling, we coded the number of 30-s intervals during which a behavior was observed at any time during that interval and divided this frequency count by the total number of 30-s intervals possible in the video of the test. To generate a proportion occurrence using instantaneous sampling, we coded behavior presence/absence at exact 30-s intervals and divided this frequency count by the total number of instant samples available for that video.

We first evaluated the appropriateness of the 43 behavior codings (BCs) for principal components analyses (PCA) by evaluating univariate distributions as well as the correlation matrix within each test (Garson, 2013). Eleven BCs suffered extreme deviations from normality (i.e., were extremely leptokurtic), eight violated assumptions of linearity (i.e., unrelated to any other BCs; correlations < 0.30), and one was redundant (i.e., almost complete covariance with one BC but not with any others; Kaiser's $MSA_{ind} = 0.267$). For predictive analyses (Section 2.3.3) we did not use the eight single BCs found to be unrelated to any others in predictive models because of our sample size, the ratio of sample size to predictors already included in models, and in order to help control study-wise Type I error rate. The 19 BCs that were extremely leptokurtic or unrelated to any other BCs captured limited information (i.e., behavior presence/absence), were rare (e.g., vocalization or urination/defecation outside designated areas), or were confounded with handler rebuke (e.g., occurrence of leash-pulling by dog was confounded/prevented by handler correcting or restraining the dog). Only the remaining 23 BCs (Table 2) are considered further.

We used data from six independent coders to provide estimates of the reliability of behavioral codings (BCs); coders were trained on coding methods prior to coding behavior for analysis. We used a fully crossed design to control for any systematic biases among coders, with six coders each independently recording data from the same ten videos per test per age (i.e., 80 videos, 20 per age, 22.7% of all videos). The six independent coders had an average of eight years of experience owning dogs (range: 0.5–15.0 years) and had owned an average of three dogs each (range: 1–7). Coders had an average of three years of scientific training (range: 0.5–6.0); additionally, two had professional experience working with dogs and three had previous experience in behavioral research. Coders were recruited via postings on numerous campus job boards across the University of Texas at Austin psychology and integrative biology departments. One coder was a psychology graduate student, one was an integrative biology undergraduate, and four were psychology undergraduates. We calculated the intraclass correlation coefficient (ICC(3,1): Hallgren, 2012; Shrout and Fleiss, 1979) for each BC to estimate the inter-rater reliability of the BC measures.

2.4. Data analysis

2.4.1. Data analysis aim 1: underlying dimensions of behavior and data reduction

To reduce the number of variables needed in subsequent analyses, we performed four separate PCAs (Tabachnick and Fidell, 1996) using SPSS 20.0.0—one for each test (ENV or S&R) and measurement type (BR or BC). Previous analyses of a larger sample of data using the same population of dogs had shown that the loadings on the PCA solution matrix for BRs were not significantly different across ages (Sinn et al., 2011); therefore, we aggregated single BCs and BRs across ages within each dog prior to each PCA here because the ratio of cases to variables was small.

We evaluated the sampling adequacy of each PCA correlation matrix using Bartlett's sphericity test and the Kaiser-Meyer-Olkin measure (Budaev, 2010; Garson, 2013). The correlation matrices of the BR and BC data in both test contexts were deemed appropriate (Tables 3–6). We determined the number of principal components to be extracted for each PCA based on a scree test, evaluation of simple and hierarchical component structure, and component interpretability (Garson, 2013; Goldberg, 2006; Zwick and Velicer, 1986). Due to potential for small sample sizes

Table 2
Operational definitions of 23 behavioral codings from the Transportation Security Administration Canine Breeding and Development Center's standardized behavioral tests. Estimates of intraclass correlation coefficients (ICC) and associated 95% confidence intervals using six independent observers are also given. ENV = environmental test; S&R = search and retrieve test; TL = test leader; TA = test assistant.

Test	Coding	Description	Reliability ICC (95% CI)
ENV	Crouch	Proportion of 30-s test intervals during which dog exhibited body-lowering postures, including crouch, crawl, or 'extended attention' (i.e., foreparts lowered, neck extended toward object)	**0.79** (0.67–0.88)
ENV	Freeze–Balk	Proportion of 30-s test intervals during which the dog interrupted dog/handler movement by freezing, balking/pulling backward against handler guidance and/or refusing to touch feet to search object, or stopping and staring	**0.96** (0.94–0.98)
ENV	Startle	Proportion of 30-s test intervals during which the dog exhibited an obvious startle response by rapidly turning and moving away from some stimulus; stimulus may not be identifiable	0.66 (0.46–0.80)
ENV	Circle–Avoid	Proportion of 30-s test intervals during which the dog exhibited avoidance of an object or task by turning its body in a roughly circular motion away from the object or task	0.65 (0.45–0.79)
ENV	Sniff Ground	Proportion of 30-s test intervals during which the dog placed its nose within ~ 6" of the surface on which it was standing (e.g., pavement, shelf, treadmill)	**0.97** (0.96–0.98)
ENV	Sniff Object	Proportion of 30-s test intervals during which the dog placed its nose within ~ 6" of an object (e.g., trashcan, floor grate, wall, door, bench) on which it was not standing	**0.96** (0.93–0.97)
ENV	Half-Hop	Proportion of 30-s test intervals during which the dog performed a 'half-hop' by placing both its front paws on an object (e.g., trashcan, cart, box); half-hop not counted if dog balks and refuses to touch paws to object	**0.99** (0.98–0.99)
ENV	Full-Hop	Proportion of 30-s test intervals during which the dog performed a 'full-hop' by jumping/leaping with all four feet onto an object or obstacle other than stationary stairs	**0.99** (0.98–0.99)
ENV	Jump TL–Toy Play	Proportion of 30-s test intervals during which dog placed its front paws on the TL during toy play or the 30-s interval after cessation of toy play	**0.95** (0.92–0.97)
ENV	Jump TL–No Toy	Proportion of 30-s test intervals during which dog placed its front paws on the TL without toy excitability	**0.97** (0.95–0.98)
S&R	Activity Level–Running	Proportion of 30-s instantaneous sample time points at which dog was running; samples with dog off camera or restrained were excluded from total	**0.90** (0.84–0.94)
S&R	Tail Position–High	Proportion of 30-s instantaneous sample time points at which dog held tail high (>5° above level); samples with tail off camera were excluded from total	**0.94** (0.91–0.97)
S&R	Tug Duration	Latency(s) for dog to release toy after initiation of post-retrieve tug session; recorded from time both TL/TA and dog grasp toy until dog releases	**1.00** (1.00–1.00)
S&R	Maximum Grip Intensity–Engaged	After tug, toy must be physically pried from dog's mouth or other extreme measures used to trigger release; maximum intensity recorded as yes/no	**0.86** (0.77–0.92)
S&R	Toy Drop Latency–Not Engaged	Latency (s) for dog to release toy during walk across field with TL & TA without stimulation/tug; walk (and latency) begins immediately when dog retrieves toy from tall grass	**0.84** (0.76–0.91)
S&R	Toy Drop Frequency	Number of times dog drops toy during entire test, from initial chase-retrieve until leaving the field; does not include brief fumble on retrieve, repositioning, or stopping to chew toy	**0.95** (0.92–0.97)
S&R	Head Turn Frequency–Hidden One	During 'hidden one' pots search exercise, number of times dog turns head >45° away from the TA and/or line of pots while toy is being hidden	**0.94** (0.91–0.97)
S&R	Search Persistence–Hidden One	During 'hidden one' pots search exercise, dog does not require assistance (e.g., step toward pots, tap pots, uncover toy) in order to persist with search and find toy; persistence recorded as yes/no	**0.93** (0.89–0.96)
S&R	Search Success Latency–Hidden One	During 'hidden one' pots search exercise, latency (s) for dog to find and grasp toy hidden under pots; search/latency begins when dog reaches the first pot and requires that the dog find the toy without accidental or intentional exposure of toy by leash or TL/TA (i.e., latency not recorded if toy exposed)	**0.77** (0.63–0.86)
S&R	Odor Indication–Hidden One	During 'hidden one' pots search exercise, dog shows clear recognition of hidden odor source by (a) a rapid reversal of movement direction toward odor source or (b) steady "bracketing" movements to hone in on odor source; odor indication recorded as yes/no	**0.75** (0.60–0.85)
S&R	Head Turn Frequency–Hidden Two	During 'hidden two' pots search exercise, number of times dog turns head >45 degrees45° away from the TA and/or line of pots while toy is being hidden	**0.93** (0.89–0.96)
S&R	Search Persistence–Hidden Two	During 'hidden two' pots search exercise, dog does not require assistance (e.g., step toward pots, tap pots, uncover toy) in order to persist with search and find toy; persistence recorded as yes/no	**0.93** (0.88–0.96)
S&R	Search Success Latency–Hidden Two	During 'hidden two' pots search exercise, latency (s) for dog to find and grasp toy hidden under pots; search/latency begins when dog reaches the first pot and requires that the dog find the toy without accidental or intentional exposure of toy by leash or TL/TA (i.e., latency not recorded if toy exposed)	**0.95** (0.92–0.97)

ICCs greater than or equal to 0.75 are shown in bold type.

Table 3
Principal component loadings of the four behavioral ratings from the TSA-CBDC standardized environment test. Ratings were averaged within individuals prior to analysis; all ratings loaded on a single component. Loadings greater than or equal to 0.50 are shown in bold type.

Behavioral Rating	Environmental Stability
Responsiveness	**0.92**
Initiative	**0.86**
Confidence	**0.81**
Concentration	**0.67**
–	
% Variance Explained	67.1
Cronbach's alpha	0.82
KMO	0.75
Bartlett's	$\chi^2_{(6)} = 89.276$, $P < 0.0001$

Bold font indicates loadings greater than or equal to 0.50.

Table 4
Principal component loadings of the five behavioral ratings from the TSA-CBDC standardized search and retrieve test. Ratings were averaged within individuals prior to analysis; all ratings loaded on a single component. Loadings greater than or equal to 0.50 are shown in bold type.

Behavioral Rating	Hunt Drive
Mental possession	**0.74**
Independent Possession	**0.66**
Hidden One	**0.64**
Hidden Two	**0.60**
Physical Possession	**0.55**
% Variance Explained	41.0
Cronbach's alpha	0.62
KMO	0.59
Bartlett's	$\chi^2_{(10)} = 39.613$, $P < 0.0001$

Bold font indicates loadings greater than or equal to 0.50.

to lead to unstable loadings, we considered only variables with a loading of at least ±0.50 (rounded to the nearest hundredth) to contribute to a component's meaning (Garson, 2013). For each PCA with more than one component, we evaluated both orthogonally and obliquely rotated solution matrices; oblique methods yielded the best simple structure for all PCAs, thus we report only the oblique solutions.

We generated age-specific unit-weighted aggregate scores for each measurement method and test based on the pattern of loadings in PCA solution matrices. We first normalized each single BR or BC variable according to its grand mean across ages; prior to normalization we removed one outlier value from the BC 'tug duration' which was 11.7 standard deviations above the mean. Resultant age-specific z values were then averaged for each dog within an age according to the pattern of loadings in the solution matrix to create a method-, age- and test-specific score. To allow scores to have the same direction of meaning (i.e., higher scores meant more of the indicated behavior), we reverse coded the loadings for the first and fourth ENV BC components and the third S&R BC component. We calculated 11 unique aggregate component scores per dog per age in this way (i.e., two aggregate scores obtained at each age using the BR method, nine aggregate scores obtained at each age using the BC method).

2.4.2. Data analysis aim 2: evaluate convergence between BR and BC methods

To test whether ratings and codings both measured the same observed behavior, we used Pearson correlations to estimate convergence between the single ENV BR aggregate score and the four ENV BC aggregate scores (four estimates), and between the single S&R BR aggregate score and the five S&R BC scores (five estimates). Aggregate scores used in correlation analysis were the average aggregate score across all four testing ages, weighted to account for varying number of observations per individual (Bland and Altman, 1995). We applied the false discovery rate (i.e., FDR = 5%; Benjamini and Hochberg, 1995) to reduce Type I error when determining statistical significance of correlations.

2.4.3. Data analysis aim 3: compare predictive validity between BRs and BCs

We used an information-theoretic approach (Burnham and Anderson 2002) and a series of mixed models to estimate repeatability and to predict binary TSA-CTES training-selection outcomes (i.e., selected vs. not selected). We first estimated repeatability/ICC for each aggregate PCA score by fitting 11 unconditional linear mixed models, each with a single PCA aggregate score as a continuous outcome variable, a constant, and individual ID and litter ID as random components. For each linear mixed model, we tested which combination of a fixed or random intercept and a fixed or random slope model provided the best fit to the data. We performed model comparisons using log-likelihood ratio tests and used maximum likelihood (ML) estimators to test for random intercepts and restricted maximum likelihood (REML) estimators to test for random slopes (Self and Liang, 1987). For each aggregate PCA score we identified the best-fit model and then extracted variance components to estimate unconditional repeatability/ICC (e.g., Boake, 1989; Dingemanse and Dochtermann, 2013; McGraw and Wong, 1996; Wolak et al., 2012). We tested whether repeatability/ICC estimates were different between the two measurement methods by converting repeatability estimates to Fisher's *z*, averaging *z*'s within a measurement method, and comparing the mean *z* values across methods with a t-test.

We had a single validity-criterion variable (i.e., selection outcome) but multiple aggregate scores/predictors measured through time; therefore we also used unconditional linear mixed models to generate best linear unbiased predictors (BLUPs) to use in subsequent generalized linear mixed models (GLMMs), which were used to predict to the TSA-CTES selection outcome. We generated two BLUPs per aggregate score from linear mixed models to represent the extent to which individual dogs were above or below the average aggregate score across the first year (the intercept BLUP) and change in aggregate score through time (the slope BLUP). We fit two unconditional linear mixed models for two different age functions with each aggregate PCA score as the outcome variable, and time (intercept only [no slope] or linear age functions), individual ID (random) and litter ID (random) as predictors. We used an information-theoretic approach (AIC values) to guide model selection among the two age functions for each mixed model (Akaike, 1981; Burnham and Anderson, 2002; Zuur et al., 2009). When we identified the most appropriate age function for each PCA aggregate score, we stopped and extracted BLUPs from that model to use as predictors of CTES selection outcomes (Croon and van Veldhoven, 2007).

To predict CTES selection outcomes, we used GLMMs on four sets of models, one for each measurement method/test context combination (i.e., BR ENV, BC ENV, BR S&R, BC S&R). Within each model set, each starting model contained an individual's age-function specific intercept and slope BLUPs (Table 7), an individual's sex, a random component predictor (i.e., litter ID), and CTES selection outcome as the dependent variable. For the BC ENV and BC S&R models, the camera type used to capture the video was also included as a fixed predictor. We were unable to include ENV TL as a fixed component predictor because only one TL handled dogs that were eventually rejected by TSA-CTES. Repeatability estimates of measurements were zero for the third and fourth ENV BC aggregate scores and the fourth S&R BC aggregate score (see Section 3.3.1); therefore, interpretation of these aggregate scores was problematic (see Section 4.2) and we dropped these aggregate scores from predictive CTES selection models.

Table 5
Principal component loadings of behavioral codings from the TSA-CBDC standardized environment test on four obliquely rotated (direct oblimin) components. Codings were averaged within individuals prior to analysis. Loadings greater than or equal to 0.50 are shown in bold type. Asterisk indicates variables that were reverse coded to maintain direction of meaning in all scores.

Behavioral Coding	Confidence*	Anxiety	Exploration	Excitability*
Full Hop	**0.91**	0.14	0.05	0.28
Circle–Avoid	−**0.70**	0.14	0.14	0.30
Freeze–Balk	−**0.66**	0.39	−0.18	0.28
Startle	−0.12	**0.82**	−0.07	−0.07
Crouch	0.23	**0.74**	0.18	0.03
Sniff Object	0.15	−0.02	**0.94**	0.13
Half Hop	−0.39	0.07	**0.70**	−0.22
Jump Handler–Toy Play	−0.10	0.09	−0.08	−**0.77**
Sniff Ground	−0.11	0.31	−0.05	**0.76**
Jump Handler–No Toy	0.02	0.44	0.06	−**0.71**
% Variance Explained	29.3	18.6	15.2	10.9
Cronbach's alpha	0.69	0.39	0.61	0.59
KMO	0.61			
Bartlett's	$\chi^2_{(45)} = 162.70$, $P < 0.0001$			

Bold font indicates loadings greater than or equal to 0.50.

Table 6
Principal component loadings of behavioral codings from the TSA-CBDC standardized search and retrieve test on five obliquely-rotated (direct oblimin) components. Codings were averaged within individuals prior to analysis. Loadings greater than or equal to 0.50 are shown in bold type; asterisk indicates variables reverse coded prior to calculation of aggregate scores to maintain direction of meaning.

Behavioral coding	Search Performance	DominantPossession	Independent Possession*	Energy Management	Search Aptitude
Search Success Latency–Hidden Two	−**0.86**	0.09	0.05	0.01	0.10
Search Persistence–Hidden Two	**0.79**	0.23	0.21	−0.11	0.18
Search Success Latency–Hidden One	−**0.56**	−0.05	0.08	−0.24	−0.45
Maximum Grip Intensity–Engaged	0.07	**0.84**	−0.09	0.03	−0.17
Tug Duration	0.13	**0.79**	−0.19	0.06	−0.03
Tail Position (High)	−0.37	**0.50**	0.28	−0.14	0.24
Toy Drop Frequency	0.15	−0.05	**0.92**	0.00	−0.14
Toy Drop Latency–Not Engaged	0.01	0.10	−**0.87**	−0.07	−0.01
Head Turn Frequency–Hidden One	0.19	−0.24	−0.06	−**0.89**	0.00
Head Turn Frequency–Hidden Two	0.07	0.33	−0.04	−**0.82**	−0.15
Activity Level (Running)	0.29	0.03	−0.06	**0.61**	−0.09
Odor Indication–Hidden One	−0.04	−0.29	−0.05	−0.07	**0.82**
Search Persistence–Hidden One	0.21	0.23	−0.14	0.20	**0.74**
% Variance Explained	24.7	15.8	12.2	11.0	9.6
Cronbach's alpha	0.50	0.09	0.25	0.60	0.56
KMO	0.58				
Bartlett's	$\chi^2_{(78)} = 217.55$, $P < 0.0001$				

Bold font indicates loadings greater than or equal to 0.50.

For each model set, our aim was to identify the most appropriate fixed structure for the GLMM model for that model set and, using that model, extract model estimates and classification outcomes (i.e., that model's predicted versus observed outcomes, expressed as percent correct). Fixed components were tested using ML estimation and Z (Diggle et al., 2002; West et al., 2006; Zuur et al., 2009). Using this GLMM approach we generated four percent-correct classifications, one for each measurement-method test-context combination. We then tested whether average percent correct was different between the BR and BC test methods using a proportions test. This proportions test represented the final indicator of predictive validity differences between the two methods.

We tested for homogeneity and normality of residuals for each GLMM model. All four GLMM models showed heterogeneity in residuals across the sexes; females had greater residual variation than males. We examined all two-way interactions between sex and random intercept and slope parameters in each of the four models; we found no significant interactions, nor was sex significant as a main effect (Section 3.3.2). Each of the four GLMMs met all other homogeneity and normality assumptions. We used SPSS 20.0 for correlation analyses and R 3.0.2 for linear mixed models and GLMMs.

3. Results

3.1. Aim 1: underlying dimensions of behavior and data reduction

3.1.1. Behavior ratings

The PCA performed on the four ENV BRs identified a single component as the best fit for the data, accounting for 67.1% of the variance in the BRs (Table 3). All of the four BRs—confidence, concentration, responsiveness, and initiative—had strong positive loadings on this single component, named here and elsewhere 'environmental stability' (Sinn et al., 2011).

The PCA performed on the five S&R BRs initially suggested a two-component solution, explaining 65.0% of the variance in the original BRs. However, subsequent predictive GLMMs fit with an aggregate score based on this second component would not converge, and unconditional repeatability/ICC estimates of one of the two aggregate scores were zero. Furthermore, previous PCA on these same ratings for a much larger sample of TSA-CBDC dogs (i.e., ~400) identified a single component named 'hunt drive' as the best fit for the data (Sinn et al., 2011). Based on this information we chose a single-component solution for the BR ratings data in S&R tests, which explained 41.0% of the variance in the ratings (Table 4).

Table 7
Model selection of age functions used to generate best linear unbiased predictors (BLUPs) for subsequent prediction. Columns provide AIC values used to select between age functions; far right column indicates which BLUP was used in generalized linear mixed models predicting training selection outcome. The third (exploration) and fourth (excitability) aggregate scores in the environment tests, as well as the fourth (energy management) aggregate score in search & retrieve tests had repeatability estimates very close (or equal to) zero, and were not included in age function analysis/subsequent prediction.

	Intercept Only (AIC)	Linear (AIC)	BLUP Age Function
Behavior ratings aggregate			
Environmental Stability	381.2	380.9	Linear
Hunt Drive	312.2	293.3	Linear
Behavior coding aggregate			
Confidence	365.9	360.2	Linear[a]
Anxiety	416.8	421.5	Intercept only
Search Performance	445.0	424.4	Linear[a]
Dominant Possession	346.4	349.0	Intercept only
Independent Possession	429.9	432.3	Intercept only
Search Aptitude	432.9	427.3	Linear[a]

[a] Individual intercept and slope BLUPs were perfectly correlated, and predictive GLMMs fitted with both parameters would not converge; only linear intercept parameters were carried forward into GLMM predictive models of TSA-CTES selection outcome. Identical GLMM predictive models that fit linear slope only models gave exact results with regards to CTES-selection.

3.1.2. Behavior codings

Four components from the ENV BC PCA were chosen as the best fit for the data and explained 73.9% of the variance in the original 10 BCs (Table 5). We chose the component names based on the operational descriptions (Table 2) of the BCs that loaded strongly on each component and on previous research (Jones and Gosling, 2005; Ley et al., 2008; Réale et al., 2007). The first component, which we named "confidence,' consisted of three BCs related to independent risk-taking behavior during ENV tests(i.e., proportion of 'froze-balked" and "circled-avoided" behaviors that were observed plus the proportion of intervals where dogs 'full-hopped'). The second component, named 'anxiety,' described variation in a tendency to crouch (Bradshaw and Nott, 1995) and display flight/startle responses. The third component, named 'exploration,' was indexed by variation in half-hopping (i.e., putting front paws up onto requested heights) and sniffing objects. The fourth component from the BCs measured in the ENV test was named 'excitability' and described variation in a dog jumping onto human testers—both with and without toy play—and distractedly sniffing odors on the ground. Component correlations among the four ENV BC components were weak; only one component correlation exceeded ± 0.20 (0.22 between confidence and excitability).

Five components from the PCA of 13 S&R BCs were selected as the best fit for the data, explaining 73.3% of the variance (Table 6). The first component, named 'search performance,' described variation in search latency and persistence in both of the pot searches. The second component, named 'dominant possession,' captured variation in tail position (Bradshaw and Nott, 1995), tug duration, and grip intensity. The third component, named 'independent possession,' described variation among individuals in their willingness to carry the toy without additional encouragement or stimulation. The fourth component, named 'energy management,' captured variation in individuals' overall energy (i.e., more or less running) but also restlessness (i.e., higher or lower 'head turn frequency' during periods of handler restraint). The fifth component, named 'search aptitude,' consisted of two BCs that described variation among individuals' performance and apperception during requests to search —'search persistence' and 'odor indications'. Component correlations among the five S&R BC components were weak; no component correlation exceeded ±0.20.

3.2. Aim 2: convergence of component scores generated by BR and BC methods

After applying the FDR (Benjamini and Hochberg, 1995), seven correlations between BC and BR scores reached statistical significance (Table 8). The ENV test BR environmental-stability score was strongly correlated with the ENV test BC confidence score ($r = 0.77$), and moderately correlated with an individual's exploration ($r = 0.34$). The S&R BR hunt-drive score was strongly correlated with the S&R BC scores for search performance and dominant possession and moderately (all r's ≈ 0.4) correlated with the three remaining S&R BC scores.

3.3. Aim 3: prediction of TSA-CTES training acceptance

3.3.1. Repeatability/ICC

The unconditional linear mixed model for the ENV test BR 'environmental stability' indicated that a random intercept was important in fitting the data (Likelihood ratio = 22.03, $P < 0.001$) but not random slopes ($\chi^2_{(3)} = 6.24$, $P = 0.10$). The unconditional repeatability/ICC estimate for environmental stability was 0.35. The mixed model for ENV BC confidence indicated that a random intercept (Likelihood ratio = 13.89, $P < 0.001$) random slope gave the best fit ($\chi^2_{(3)} = 11.70$, $P = 0.008$). The unconditional repeatability/ICC estimate for confidence was 0.26. The mixed model for ENV BC anxiety indicated that a random intercept (Likelihood ratio = 6.13, $P = 0.01$) fixed slopes ($\chi^2_{(3)} = 1.24$, $P = 0.74$) model best fit the data. The unconditional repeatability/ICC estimate for anxiety was 0.16. The mixed model for ENV BC exploration indicated that a fixed intercept (Likelihood ratio = 7.1×10^{-8}, $P = 0.9998$) and fixed slopes ($\chi^2_{(3)} = 0.76$, $P = 0.86$) model provided the best fit. The unconditional repeatability/ICC estimate for exploration was 0.00. The mixed model for ENV BC excitability indicated that a fixed intercept (Likelihood ratio = 2.67, $P = 0.10$) random slopes ($\chi^2_{(3)} = 15.88$, $P = 0.001$) model provided the best fit. The repeatability/ICC estimate for excitability was 0.09.

The mixed model for S&R BR hunt drive indicated that a random intercept (Likelihood ratio = 22.45, $P < 0.001$) random slopes ($\chi^2_{(3)} = 24.86$, $P < 0.001$) model best fit the data. The unconditional repeatability/ICC estimate for hunt drive was 0.28. The mixed model for S&R BC search performance indicated that a random intercept (Likelihood ratio = 5.58, $P = 0.02$) random slopes ($\chi^2_{(3)} = 26.59$, $P < 0.001$) model provided the best fit. The repeatability/ICC estimate for search performance from this model was 0.11.The mixed model for S&R BC dominant possession indicated that a random intercept (Likelihood ratio = 32.16, $P < 0.001$) fixed slopes ($\chi^2_{(3)} = 3.39$, $P = 0.33$) model provided the best fit. The repeatability/ICC estimate for S&R BC dominant possession from this model was 0.16. The mixed model for S&R BC independent possession indicated that a random intercept (Likelihood ratio = 32.98, $P < 0.001$) fixed slopes ($\chi^2_{(3)} = 3.59$, $P = 0.31$) model provided the best fit. The repeatability/ICC estimate for independent possession from this model was 0.32. The mixed model for S&R BC energy management indicated that a random intercept (Likelihood ratio = 21.55, $P < 0.001$) random slopes ($\chi^2_{(3)} = 10.80$, $P = 0.01$) model provided the best fit. The repeatability/ICC estimate for energy management from this model was 0.09. The mixed model for S&R BC search aptitude indicated that a random intercept (Likelihood ratio = 6.47, $P = 0.01$) random slopes ($\chi^2_{(3)} = 11.61$, $P = 0.001$) model provided the best fit. The unconditional repeatability/ICC estimate for search aptitude from this model was 0.13.

The average repeatability detected by BR methods was 0.31 and the average repeatability detected by BC methods was 0.15; there were borderline statistically significant differences in the average

Table 8
Pearson correlations among weighted averages (across ages) of behavioral rating (BR) components and behavioral coding (BC) components (all N = 52). Bold font indicates results that are statistically significant after implementation of the false discovery rate. ENV = environmental test; S&R = search and retrieve test.

BC Aggregate Score		BR Aggregate Score	
		ENV	S&R
ENV	Confidence	**0.77** ($P < 0.001$)	
	Anxiety	−0.26 ($P = 0.068$)	
	Exploration	**0.34** ($P = 0.015$)	
	Excitability	0.13 ($P = 0.358$)	
S&R	Search Performance		**0.57** ($P < 0.001$)
	Dominant Possession		**0.47** ($P < 0.001$)
	Independent Possession		**0.42** ($P < 0.001$)
	Energy Management		**0.44** ($P < 0.001$)
	Search Aptitude		**0.37** ($P < 0.001$)

repeatability estimate between the two methods (paired t-test: $t_{(1)} = 8.37$, P = 0.08).

3.3.2. Predictive validity

The four final models for each model set are given in Table 9. For the ENV BR method, the final predictive model had fixed components that were all $P < 0.05$, with the exception of an individual dog's sex. An individual's environmental stability—both its behavior and behavioral change relative to the average behavior and change—were significant predictors of CTES selection outcomes. Using the BC methods on ENV test results indicated that using a headcam and an individual's positive change in confidence relative to the average change were significant. In the final S&R BR model, positive linear developmental change of hunt drive was important to CTES selection, but not the dogs' behavior across the first year of life relative to the average. In the final BC S&R model, using a handheld camera rather than a headcam improved chances of selection, as did an individual's average dominant possession, but dog sex, as well as search performance, independent possession, and search aptitude did not impact selection (Fig. 1).

The correct classification rate for the ENV BR variable was 84.6%, 88.5% for the ENV BC variables, 88.5% for the S&R BR variable, and 86.5% for the S&R BC variables. There was no difference in classification rates between the BC and BR methods ($\chi_{(1)}{}^2 = 0.00$, P = 1.00).

4. Discussion

A primary goal of research on working-dog populations is to use behavioral assessments to predict their subsequent working potential. However, most behavioral assessments use only a single measurement method (but see Mirkó et al., 2013; Wilsson and Sinn 2012), despite measurement theory that predicts that different measurement methods may capture different aspects of behavior (Freeman et al., 2011; Furr and Funder 2007). These measurement differences may result in differences in predictive validity for 'success' outcomes in working dogs; however, this idea is rarely tested.

In the service of improving measurement practices in animal behavior research, the present study compares rating and coding approaches for the same group of dogs during the same behavior tests. We demonstrated that high inter-observer reliability characterized the codings methods used to measure behavior in TSA-CBDC standardized tests; previous research had already established the existence of high inter-observer reliability for the ratings method used here (Fratkin et al., 2015). Some researchers may be hesitant to employ ratings methods in order to avoid the observer biases widely thought to be associated with such methods. However, our results add to the research literature indicating that ratings methods may be just as reliable (i.e., have high inter-observer agreement) as the supposedly more 'objective' codings methods (Fratkin et al., 2013; Vazire et al., 2007).

In terms of practical implications, this study is consistent with our previous research Fratkin et al. (2015) demonstrating that relative novices can be trained to perform codings and ratings with high levels of inter-observer repeatability (on par with veteran dog handlers). However, the time required to perform the two methods differed substantially. For the environmental test, ratings methods (including the time to perform the assessment) required approximately10–30 min, whereas recording codings from video required 20–60 min. Similarly, in the search & retrieve test, rating methods required 6–15 min, while codings methods tripled this time requirement. Both rating and coding methods involved some basic quality-control review as well as data entry; however the time requirements for these were similar across the two methods.

Statistical aggregation (PCA) of codings and ratings indicated that coding methods tended to capture finer-grained information about between-individual differences in behavior (multiple aggregate descriptors), whereas all single ratings tended to be tapping the same underlying factor in each standardized test (i.e., the PCA yielded a single aggregate dimension). For the S&R test, strong convergence between the multiple coding aggregate scores and the single rating aggregate score indicated that both measurement methods also tended to measure the same underlying construct. In ENV tests, however, two aggregate traits emerged from PCA of the codings (anxiety and excitability) that were not captured by the single rating aggregate, indicating that, in some situations, one measurement method may reliably capture behavior that is not captured by another (Carter et al., 2012).

Use of ratings methods tended to capture behavior that was more repeatable across time relative to codings methods. Hunt drive and environmental stability (aggregate scores measured by rating methods) yielded significant between-individual differences in behavior (repeatability ~0.31). Aggregate scores generated by coding methods, on the other hand, appeared to capture behaviors that were characterized by greater amounts of within-individual variation across time (i.e., were less predictable, average repeatability = 0.15). Indeed, several aggregate scores yielded by coding methods were not repeatable at all (i.e., exploration and excitability in ENV tests) or had very low repeatability estimates (i.e., energy management, search performance, and search aptitude in S&R tests; anxiety in ENV tests).

Both measurement methods resulted in high rates of predictive validity; we detected no differences between measurement methods in their ability to correctly classify training-selection outcome. Below, we discuss in more detail each of these major findings and their implications for working-dog development and breeding programs.

Table 9
Final model estimates for the generalized linear mixed models used in prediction of TSA-CTES training acceptance. The comparison group for 'female' is male, and for 'headcam' is 'hand-held camera'. The estimate for Litter ID is its variance component, and the standard error is its standard deviation. BR = behavior rating; BC = behavior coding; ENV = environment test; S&R = search and retrieve test; Intercept.linear = intercept BLUP; test.age.linear = slope BLUP; Intercept.int = intercept-only BLUP.

Parameter	Estimate	Std Error	Z	*P*
Environment test behavioral ratings				
(Intercept)	2.64	0.78	3.36	0.001
litter ID	1.2×10^{-10}	1.1×10^{-5}		
female	−1.42	0.96	−1.49	0.13
BR_ENV_environmental stability.Intercept.linear	1.66	0.74	2.24	0.02
BR_ENV_environmental stability.test age.linear	1.71	0.73	2.35	0.02
Environment test behavioral codings				
(Intercept)	4.13	1.25	3.29	0.001
litter_ID	0.25	0.50		
female	−1.92	1.19	−1.61	0.11
headcam	−3.20	1.51	−2.12	0.03
BC_ENV_confidence.Intercept.linear	1.24	0.52	2.40	0.02
BC_ENV_anxiety_Intercept.int	0.50	0.54	0.93	0.35
Search and retrieve test behavioral ratings				
(Intercept)	2.45	0.84	2.91	0.004
litter_ID	1.23	1.11		
female	−0.43	0.96	−0.44	0.66
BR_S&R_hunt drive.Intercept.linear	0.58	0.49	1.19	0.23
BR_S&R_hunt drive.test age.linear	1.52	0.60	2.53	0.01
Search and retrieve test behavioral codings				
(Intercept)	3.60	1.11	3.25	0.001
litter_ID	0.00	0.00		
female	−0.54	0.96	−0.56	0.57
headcam	−3.10	1.51	−2.05	0.04
BC S&R search performance.Intercept.linear	−0.30	0.64	−0.47	0.64
BC_S&R_dominant possession.Intercept.int	1.41	0.66	2.13	0.03
BC S&R independent possession.Intercept.int	0.54	0.52	1.05	0.29
BC S&R search aptitude.Intercept.linear	−0.06	0.60	−0.10	0.92

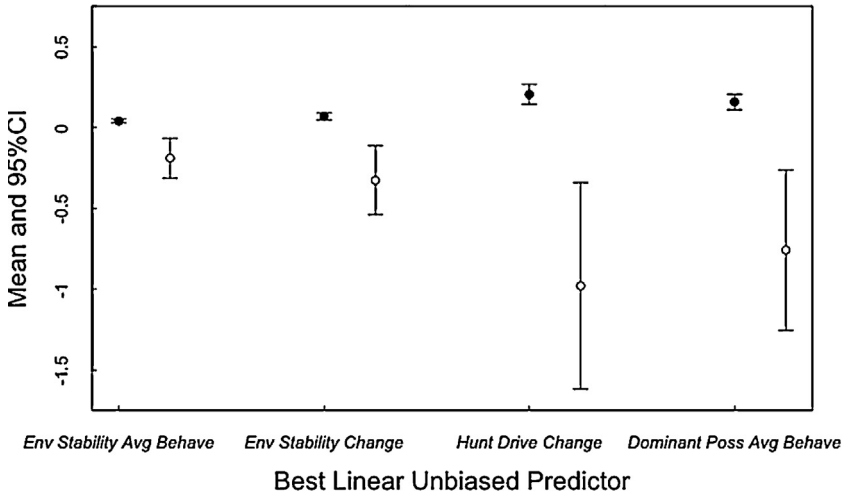

Fig. 1. Mean best-linear unbiased parameter (BLUP) values were generated from mixed models using PCA scores of ratings and codings as predictors of selection for training. The BLUPs for ratings and codings that were significant in predictive models are displayed; black circles = selected dogs; white circles = non-selected dogs. Intercept BLUPs (labeled 'avg behave') are the difference between individual and group mean PCA scores and can be conceptualized as an individual's relative distance from the observed mean. Slope BLUPs (labeled 'change') are the difference between the individual's slope (regression on PCA scores through time) and mean slope for the sample and thus indicate individual change relative to the mean change for the group. Significant BLUPs shown in the figure were derived from environmental stability ratings from environmental tests (ENV) and hunt drive ratings and dominant possession codings from search & retrieve tests (S&R). One single coding from ENV tests ('confidence') was strongly correlated (0.8) with the rating 'environmental stability' given during the same tests. Predictive models and BLUP graphs for confidence match those given for environmental stability below.

4.1. Repeatability of aggregate scores

Statistical aggregation of ratings in each standardized test resulted in two measures that were moderately repeatable; that is, using a ratings method one could characterize dogs as having moderately predictable behavior from one time period to the next. In addition, estimates of between- and within-individual variation indicated that environmental stability (ENV test) for all individuals in our sample tended to change in the same way through time (i.e., a fixed slope unconditional repeatability model, where each individual regression line had the same slope, was the best-fit model). This was not the case for hunt drive (S&R test), which tended to change in a different way across individuals in our sample (i.e., a random slope model was the best-fit for these data).

Statistical aggregation of codings resulted in a four measures that had an overall average repeatability that was not statisti-

cally different from the overall average repeatability estimates for ratings. However, use of two of the coding aggregates in mixed models indicated that these aggregate scores did not reflect consistent between-individual differences in behavior (i.e., exploration and excitability from ENV tests resulted in best-fit fixed intercept repeatability models, and estimates of repeatability not different from zero). In addition, five coding aggregate scores (i.e., anxiety in ENV tests, search performance, dominant possession, energy management, and search aptitude in S&R tests) had low repeatability estimates (<0.20), indicating that within-individual variance was high relative to total variance in the sample. In other words, while PCA analysis partitioned between-individual variation into several different components within each test using coding methods, aggregate codings also appeared to capture high and unpredictable within-individual variation in behavior through time relative to ratings methods (Dingemanse and Dochtermann, 2013).

We also observed strong correlations between codings and ratings in one test (S&R). Taken together, our repeatability and convergence results highlight the idea that even when measuring the same underlying construct in the same test using different measurement methods (e.g., in S&R tests), measurements with narrower measurement breadth (i.e., codings) may also be more transient or situation-specific. This conclusion differs from the results of a recent meta-analysis, which suggested that on average, codings methods and ratings methods tend to yield similar levels of repeatability (Fratkin et al., 2013). Further work is necessary in order to understand how measurement breadth and repeatability may or may not covary depending on the situations in which testing occurs. This issue is important because ratings methods typically involve less effort that coding methods do; in working-dog standardized tests used by programs such as the TSA-CBDC, the use of the current ratings system or even an expanded one could be based on pragmatic concerns (i.e., the relatively small effort needed to use a rating method rather than a coding approach). Of course, predictive validity may be the ultimate criterion to consider when adjudicating between the two methods.

4.2. Relative predictive validity of BR and BC methods

The primary aim of this study was to test whether two different measurement methods would yield differences in predictive validity with respect to training-selection outcomes. We found that both BC and BR model sets had high correct classification rates and there were no differences detected in predictive validity between the two methods.

All four predictive model sets (i.e., BR ENV, BC ENV, BR S&R, BC S&R) resulted in final models with high percentages of correct classifications. Across the four model sets, we also observed no sex-specific selection bias (female and male dogs were equally likely to be selected for training). Each model classified individual dogs as 'not selected', both correctly and incorrectly, which was reassuring given the heavily unbalanced validity-criterion variable (i.e., 83% of dogs in our initial sample were 'selected'). In the ENV test, an individual's average environmental stability, as measured by the ratings method, as well as positive linear change in environmental stability were important predictors of CTES selection, as was average confidence as measured by the coding method. In the S&R test, positive linear change in hunt drive over the first year of life (but not average hunt drive) as measured by ratings was an important predictor of selection, as was an individual's dominant possession over its first year of life as measured by codings.

In practical terms, this finding suggests that attempts to encourage environmental stability (as measured by ratings methods) and confidence (as measured by coding methods) over the first year of life would improve selection outcomes. It is worth noting that confidence, measured by codings, and environmental stability, measured by ratings, appeared to be essentially equivalent ($r = 0.77$). In S&R tests, selection outcomes could be improved through focus on an individual's dominant possession scores (coding methods) and encouraging positive change in hunt drive over time (ratings methods). Interestingly, only change in hunt drive, and not the average level of an individual's behavior mattered to training-selection success, suggesting that tracking whether an individual is increasing their hunt drive over time (rather than focusing on how an individual behaves at any one given time) would be recommended.

4.3. Limitations and directions for future research

Several caveats to the present study must be mentioned. First, the ratings methods used were limited, in the sense that they were part of an already existing TSA-CBDC standard operating procedure. Evaluation and generation of a wider range of operational definitions for reliable behavior that could be measured by BR methods may result in greater predictive validity for the BR method. Our finding that two aggregate traits–anxiety and excitability—that emerged from PCA of the coding methods were not captured by the single rating also suggests that systematic generation of novel ratings to use for measuring more behavior in standardized tests would be warranted (see also Uher and Asendorpf, 2008; Wilsson and Sinn 2012).

Second, we used a single dataset both to generate the model parameters and to predict a success criteria; a more stringent test would be to quantify classification rates for the model parameters generated from this data set on an independent one (e.g., Sinn et al., 2010; Wilsson and Sinn, 2012). Also, because our initial sample was inevitably unbalanced (83% of the dogs were selected for training), a 'ceiling' effect on classification rates may have been reached. Codings and ratings may have had differences in predictive validity but the differences did not emerge because there was not much room for improvement.

Third, we used selection for training as our measure of success but there are several other reasonable criteria for success. Any dog that is ultimately successful must successfully qualify for training, but quantification of different aspects of success, such as training success (i.e., days spent in training, rate of learning new odors, etc.) or real-life working outcomes (i.e., number of annual certifications) is needed to augment our findings. Each potential criterion has its own set of statistical, conceptual, and methodological pros and cons. The main disadvantage of using acceptance into training or certification is that these criteria might not predict ultimate success in the field. Major disadvantages of criteria further down the chain (e.g., field success) are the reduced sample sizes available and the severe restrictions of range in behavior that result from the systematic exclusion of dogs at each progressive stage; the restriction of range is a major problem because it could result in variables that are genuinely predictive of the outcome appearing not to be predictive. More pragmatically, measures of field success may simply not be available; after dogs have been in the field for long enough to demonstrate their performance, they are institutionally and geographically dispersed, making systematic, standardized measures of success difficult to gather. Even if measures of field success were available, another major concern is that such measures may have poor construct validity.

After spending significant time in the field, numerous other factors (beyond the dog's behavioral tendencies) are likely to affect success; such factors include differences in amount of time spent in the field, differences in opportunities to demonstrate success, differences in experience (e.g., exposure to health or psychological risks) and differences between different handlers' skills, experience levels, and personalities. These differences are overlaid with individuals' and organizations' incentives to present favorable

characterizations of their dogs' performance. This is not to say that measures of field success are poor outcome measures; it just means that, like all measures, they have limitations and are not the clearly optimal measure they are sometimes assumed to be. More broadly, research on differences in predictive validity between measurement methods is in its infancy with regards to animal personality in general, and for working dogs in particular. More work on the most valid ways of measuring animal behavior is needed.

It is worth noting however that evidence is emerging that, in many cases, the use of different measurement methods may not impact estimates of behavioral repeatability in dogs (Fratkin et al., 2013). There are also preliminary indications that the breadth of the construct measured may not impact the reliability and validity of the measurements. In one study BR methods with different measurement breadth (behavioral ratings similar to the ones used here as well as broader subjective ratings that were more adjective-based, less dependent on situations, and based more on observer intuition) yielded similar predictive success with respect to outcomes in a working-dog program (Wilsson and Sinn 2012). In principle rating methods could be used with regard to even narrower constructs (i.e., less "distance" from specific behaviors) than the BRs used in the present research. More research is needed to evaluate the impact of aggregation level on predictive validity.

4.4. Conclusions

Predicting success in working roles from behavioral assessments made early in life is perhaps the 'gold standard' of most working-dog programs. Unfortunately, little is known concerning how best to quantify behavior during assessments and even prediction of whether a purpose-bred dog will be selected for training often proves difficult. Measurement theory suggests that different measurement approaches (i.e., codings versus ratings) may have important differences with regards to predictive validity because the two methods can capture different aspects of behavior. However, our findings suggest that these theoretical differences between the methods may not play out in practice, even when the different methods result in quantification of different components of variation between individuals (see also Wilsson and Sinn 2012). Ratings methods can be less time consuming (especially if performed by the handler immediately after the test) so for programs with limited time, the trait-rating measures may be preferable to the more labor-intensive coding methods. It is worth noting, however, that with minimal training, novices can undertake behavioral codings and ratings that are reliable and converge with the scores of experts (Fratkin et al., 2015); therefore, the time savings afforded by ratings may not be as critical as they would have been had expert labor been required. Nonetheless, more work on the different measurement methods is needed to address some key outstanding questions: Do choices about measurement method matter to some types of outcome criteria but not others? What aspects of standardized test situations result in cases where codings and ratings capture the same or different aspects of observed behavioral variation? Given the costs of rearing, caring for, and training working dogs, these questions are consequential for improving the efficiency of working-dog programs.

In addition, we note that the two studies to date that directly address this question in working dogs (the present work; Wilsson and Sinn 2012) compare different measurement methods, but both studies based each comparison on standardized test situations. Thus, it is possible that different measurement methods may have yielded differences in predictive validity in other measurement scenarios, including less structured observations or observations in seemingly unrelated settings. Recent work suggests that off-duty behavior may be an important component in predicting 'real-life' working success in odor-detection dogs (Rocznick et al., 2015). Thus, research may need to expand its scope with regards to situations during early life in which behavior is measured, as well as continue to assess which measurement methods may be most applicable to different situations (McGarrity et al., 2015). Clearly, both between- and within-individual variation in behavior is ubiquitous in working-dog populations; understanding the best ways of quantifying this variation remains an outstanding issue in animal personality research in general, and for working dogs in particular.

Conflict of interest

None.

Acknowledgments

Aleyda Vizzueth Herrera, Isaac Miller-Crews, Ryan Milton, Kristen Murrieta, and Jessica Tauber coded behavior from video for estimation of reliability. Miles Bensky, Stephen DeBono, Jamie Fratkin, and Katharine Lee assisted with video recording. Research was conducted in accordance with University of Texas Institutional Animal Care and Use Protocol # AUP-2010-00061. Funding was provided by the US Department of Homeland Security, Science & Technology Directorate, contract HSHQDC-10-C-00085: "Improving the Effectiveness of Detector-Dog Selection and Training through Measurement of Behavior and Temperament".

References

Akaike, H., 1981. Likelihood of a model and information criteria. J. Econ. 16, 3–14.
Batt, L.S., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2008. Factors associated with success in guide dog training. J. Vet. Behav. 3, 143–151.
Beaudet, R., Chalifoux, A., Dallaire, A., 1994. Predictive value of activity level and behavioral evaluation on future dominance in puppies. Appl. Anim. Behav. Sci. 40, 273–284.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B. 57, 289–300.
Bland, J.M., Altman, D.G., 1995. Calculating correlation coefficients with repeated observations: part 2—correlation between subjects. Brit. Med. J. 310, 633.
Boake, C., 1989. Repeatability: its role in evolutionary studies of mating behavior. Evol. Ecol. 3, 173–182.
Bradshaw, J.W.S., Nott, H.M.R., 1995. Social and communication behaviour of companion dogs. In: Serpell, J. (Ed.), The Domestic Dog: Its Evolution, Behaviour, and Interactions with People. Cambridge University Press, Cambridge, pp. 115–130.
Budaev, S.V., 2010. Using principal components and factor analysis in animal behavior research: caveats and guidelines. Ethology 116, 472–480.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodal Inference. Springer, New York (488 pp.).
Campbell, W.E., 1975. Behaviour Problems in Dogs. American Veterinary Publications, California.
Capitanio, J.P., 1999. Personality dimensions in adult male rhesus macaques: prediction of behaviors across time and situation. Am. J. Primatol. 47, 299–320.
Carter, A.J., Marshall, H.H., Heinsohn, R., Cowlishaw, G., 2012. Evaluating animal personalities: do observer assessments and experimental tests measure the same thing? Behav. Ecol. Sociobiol. 66, 153–160.
Champness, K.A., 2000. Revised Breeding and Rearing Strategy for FAA Detector Dogs. Federal Aviation Administration, San Antonio, Texas.
Champness, K.A., 1996. Development of a Breeding Program for Drug Detector Dogs: Based on Studies of a Breeding Population of Guide Dogs, PhD Thesis. Department of Agriculture and Resource Management, The University of Melbourne.
Croon, M.A., van Veldhoven, M.J.P.M., 2007. Predicting group-level outcome variables from variables measured at the individual level: a latent variable multilevel model. Psychol. Methods 12, 45–57.
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.
Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L., 2002. The Analysis of Longitudinal Data, 2nd edition. Oxford University Press, Oxford, England.
Dingemanse, N.J., Dochtermann, N.A., 2013. Quantifying individual variation in behaviour: mixed-effect modelling approaches. J. Anim. Ecol. 82, 39–54.
Duffy, D.L., Serpell, J.A., 2008. Behavioral assessment of guide and service dogs. J. Vet. Behav. Clin. Appl. Res. 3, 186–188.
Duffy, D.L., Serpell, J.A., 2012. Predictive validity of a method for evaluating temperament in young guide and service dogs. Appl. Anim. Behav. Sci. 138, 9–109.

Fratkin, J.L., Sinn, D.L., Patall, E.A., Gosling, S.D., 2013. Personality consistency in dogs: a meta-analysis. PLoS One 8, e54907.
Fratkin, J.L., Sinn, D.L., Thomas, S., Hilliard, S., Olson, Z., Gosling, S.D., 2015. Do you see what I see? Can non-experts with minimal training reproduce expert ratings in behavioral assessments of working dogs? Behav. Processes 110, 105–116.
Freeman, H., Gosling, S.D., Schapiro, S.J., 2011. Comparison of methods for assessing personality in nonhuman primates. In: Weiss, A., King, J.L., Murray, L. (Eds.), Personality and Temperament in Nonhuman Primates, Developments in Primatology: Progress and Prospects. Springer Science, New York, pp. 17–40.
Furr, R.M., Funder, D.C., 2007. Behavioral observation. In: Robins, R.W., Fraley, R.C., Krueger, R.F. (Eds.), Handbook of Research Methods in Personality Psychology. Guilford Press, New York, pp. 273–291.
Garson, G.D., 2013. Factor Analysis. Statistical Associates Publishers, Asheboro, NC.
Gartner, M.C., Powell, D., 2012. Personality assessment in snow leopards (*Uncia uncia*). Zoo Biol. 31, 151–165.
Goddard, M.E., Beilharz, R.E., 1982. Genetic and environmental factors affecting the suitability of dogs as guide dogs for the blind. Theor. Appl. Genet. 62, 97–102.
Goddard, M., Beilharz, R., 1986. Early prediction of adult behavior in potential guide dogs. Appl. Anim. Behav. Sci. 15, 247–260.
Goldberg, L.R., 2006. Doing it all bass-ackwards: the development of hierarchical factor structures from the top down. J. Res. Pers. 40, 347–358.
Gosling, S.D., 2001. From mice to men: what can we learn about personality from animal research? Psychol. Bull. 127, 45–86.
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.
Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. 8, 23–34.
Helton, W.S., 2009. Canine Ergonomics: the Science of Working Dogs. CRC Press, New York.
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.
Highfill, L., Hanbury, D., Kristiansen, R., Kuczaj, S., Watson, S., 2010. Rating vs coding in animal personality research. Zoo Biol. 29, 509–516.
Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J. Am. Vet. Med. Assoc. 223, 1293–1300.
Jones, A.C., Gosling, S.D., 2005. Temperament and personality in dogs (*Canis familiaris*): a review and evaluation of past research. Appl. Anim. Behav. Sci. 95, 1–53.
King, T., Marston, L.C., Bennett, P.C., 2012. Breeding dogs for beauty and behaviour: why scientists need to do more to develop valid and reliable behaviour assessments for dogs kept as companions. Appl. Anim. Behav. Sci. 137, 1–12.
Kubinyi, E., Gosling, S.D., Miklósi, Á., 2015. A comparison of rating and coding behavioural traits in dogs. Acta Biol. Hung. 66, 27–40.
Ley, J., Bennett, P., Coleman, G., 2008. Personality dimensions that emerge in companion canines. Appl. Anim. Behav. Sci. 110, 305–317.
Lloyd, A.S., Martin, J.E., Bornett-Gauci, H.L.I., Wilkinson, R.G., 2007. Evaluation of a novel method of horse personality assessment: rater-agreement and links to behaviour. Appl. Anim. Behav. Sci. 105, 205–222.
Maejima, M., Inoue-Murayama, M., Tonosaki, K., Matsuura, N., Kato, S., Saito, Y., Weiss, A., Murayama, Y., Ito, S.I., 2007. Traits and genotypes may predict the successful training of drug detection dogs. Appl. Anim. Behav. Sci. 107, 287–298.
Martin, P., Bateson, P., 1993. Measuring Behaviour: An Introductory Guide, 2nd ed. Cambridge University Press, Cambridge.
McGarrity, M.E., Sinn, D.L., Gosling, S.D., 2012. Supplemental report 2: descriptive analysis of behavioral assessment data from TSA-CBDC stimulation sessions and TSA-CTES drive building exercises and final evaluations. US Department of Homeland Security, Science & Technology Directorate, Contract HSHQDC-10-C-00085: Improving the Effectiveness of Detector-Dog Selection and Training through Measurement of Behavior and Temperament. 50pp.
McGarrity, M.E., Sinn, D.L., Gosling, S.D., 2015. Which personality dimensions do puppy tests measure? A systematic procedure for categorizing behavioral assays. Behav. Proc. 110, 117–124, http://dx.doi.org/10.1016/j.beproc.2014.09.029.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46.
Mirkó, E., Dóka, A., Miklósi, A., 2013. Association between subjective rating and behaviour coding and the role of experience in making video assessments on the personality of the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci., http://dx.doi.org/10.1016/j.applanim.2013.10.003. Netto, W.J., Planta, D.J.U., 1997. Behavioral testing for aggression in the domestic dog. Appl. Anim. Behav. Sci. 52, 243–263.
Réale, D., Reader, S.M., Sol, D., McDougall, P.T., Dingemanse, N.J., 2007. Integrating animal temperament within ecology and evolution. Biol. Rev. 2007, 291–318.
Rocznick, D., Sinn, D.L., Thomas, S., Gosling, S.D., 2015. Criterion analysis and content validity for standardized behavioral tests in a detector-dog breeding program. J. Forensic Sci. 60, S213–S221.
Rooney, N.J., Bradshaw, J.W.S., Almey, H., 2004. Attributes of specialist search dogs—a questionnaire survey of UK dog handlers and trainers. J. Forensic Sci. 49, 300–306.Self, S.G., Liang, K.Y., 1987. Asymptomotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82, 605–610.
Serpell, J.A., Hsu, Y.Y., 2001. Development and validation of a novel method for evaluating behavior and temperament in guide dogs. Appl. Anim. Behav. Sci. 72, 347–364.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
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.
Sinn, D.L., Hixon, G., Gosling, S.D., 2011. Tasks 2.2, 2.3, and 3.2 combined Report—exploratory factor analysis, internal validity of aggregate behavior scales, and test-retest correlations in the TSA-CBDC behavioral test data. US Department of Homeland Security, Science & Technology Directorate, Contract HSHQDC-10-C-00085: Improving the Effectiveness of Detector-Dog Selection and Training through Measurement of Behavior and Temperament. 72pgs.
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., Forkman, B., 2002. Personality traits in the domestic dog (*Canis familiaris*). Appl. Anim. Behav. Sci 79, 133–155.
Svartberg, K., 2002. Shyness–boldness predicts performance in working dogs. Appl. Anim. Behav. Sci. 79, 157–174.
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.
Tabachnick, B.G., Fidell, L.S., 1996. Using Multivariate Statistics, 3rd ed. Harper Collins, New York.
Tomkins, L.M., Thomson, P.C., McGreevy, P.D., 2011. Behavioral and physiological predictors of guide dog success. J. Vet. Behav. Clin. Appl. Res. 6, 178–187.
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. Guilford, Guilford, New York, New York, pp. 190–206.
West, B., Welch, K.B., Galecki, A.T., 2006. Linear Mixed Models: A Practical Guide Using Statistical Software. Chapman and Hall/CRC Press.
Wilsson, E., Sinn, D.L., 2012. Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program. Appl. Anim. Behav. Sci. 141, 158–172.
Wilsson, E., Sundgren, P.E., 1997. The use of a behavior test for the selection of dogs for service and breeding. 1. Method of testing and evaluating test results in the adult dog, demands on different kinds of service dogs, sex and breed differences. Appl. Anim. Behav. Sci. 53, 279–295.
Wilsson, E., Sundgren, P.E., 1998. Behavior test for eight-week-old puppies: heritability's of tested behavior traits and its correspondence to later behaviour. Appl. Anim. Behav. Sci. 58, 151–162.
Wolak, M.E., Fairbairn, D.J., Paulsen, Y.R., 2012. Guidelines for estimating repeatability. Methods Ecol. Evol. 3, 129–137.
Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, New York, pp. 574.
Zwick, W.R., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. Psychol. Bull. 99, 432–442.
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.