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*

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.

© 2016 Elsevier B.V. All rights reserved.

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).

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