Received: 16 February 2018 | Revised: 4 January 2019 | Accepted: 11 January 2019

DOI: 10.1111/eth.12846

RESEARCH PAPER

WILEY ethology

What are we measuring? Novices agree amongst themselves (but not always with experts) in their assessment of dog behaviour

Kirke L. Munch[1] | Erik Wapstra[1] | Scott Thomas[2] | Michelle Fisher[3] | David L. Sinn[1,4,5]

[1]School of Biological Sciences, University of Tasmania, Hobart, Tasmania, Australia

[2]American Kennel Club's Detection Dog Task Force, Castroville, Texas

[3]Australian Border Force Detector Dog Program, Australian Custom and Border Protection Service, Melbourne, Victoria, Australia

[4]Department of Psychology, University of Texas at Austin, Austin, Texas

[5]Environmental Science and Policy, University of California, Davis, California

Correspondence
Kirke L. Munch, School of Biological Sciences, University of Tasmania, Hobart, Tas., Australia.
Email: klmunch@utas.edu.au

Funding information
US Department of Homeland Security, Grant/Award Number: HSHQDC-10-C-00085; Australian Research Council

Abstract

Humans differ in how they perceive, assess, and measure animal behaviour. This is problematic because strong observer bias can reduce statistical power, accuracy of scientific inference, and in the worst cases, lead to spurious results. Unfortunately, reports and studies of measurement reliability in animal behaviour studies are rare. Here, we investigated two aspects of measurement reliability in working dogs: inter-observer agreement and criterion validity (comparing novice ratings with those given by experts). Here, we extend for the first time a powerful framework used in human psychological studies to investigate three potential aspects of (dis)agreement in non-human animal behaviour research: (a) that some behaviours are easier to observe than others; (b) that some subjects are easier to observe than others; and (c) that observers with different levels of experience with the subject animal give the same or different ratings. We found that novice observers with the same level of experience agreed upon measures of a wide range of behaviours. We found no evidence that age of the dogs affected agreement between these same novice observers. However, when observers with different levels of experience (i.e., novices vs. a working dog expert) assessed the same dogs, agreement appeared to be strongly affected by the measurement instrument used to assess behaviour. Given that animal behaviour research often utilizes different observers with different levels of experience, our results suggest that further tests of how different observers may measure behaviour in different ways are needed across a wider variety of organisms and measurement instruments.

KEYWORDS

agreement, *Canis familiaris*, observer bias, personality, ratings

1 | INTRODUCTION

A fundamental issue in scientific research is that human observations are biased (Meagher, 2009). The idea that different people may give different assessments of the same observation or phenomena is especially relevant in studies of animal behaviour, where function, motivation, or purpose of the nonhuman subject animal (or lack thereof) is often implied. Added to this inherent human bias are two problematic issues: one, it is typical often only to have a single observer rate or code observed behaviours of interest in studies of animal behaviour (reviewed in Burghardt et al., 2012; Kaufman & Rosenthal, 2009), and two, human disagreement of the definition of behaviour is widespread. In a survey of members of three scientific societies involved in animal behaviour research, Levitis, Lidicker, and Freund (2009) found that respondents, when asked to define "behaviour", tended to contradict themselves, each other and published definitions, leading the authors to conclude that humans involved in animal behaviour research use individually variable intuitive (and therefore subjective) meanings of common, fundamental concepts in the field. Simply put, human measures of animal behaviour can reflect properties of the human observer (and not of the subject animal), but the extent to which this is the case is currently unknown.

The impact of human bias on scientific inference in animal behaviour studies has a long history dating from the 1970s (Caro, Roper, Young, & Dank, 1979; Johnson & Bolstad, 1973), and sections on reliability of measurement are a regular component of many animal behaviour textbooks (e.g., Martin & Bateson, 1993). Two reviews, however, point to the general lack of reliability controls in recent animal behaviour research. Kaufman and Rosenthal (2009) reviewed 100 articles from two volumes of Animal Behaviour and found that 96 of these studies failed to report any estimates of inter-observer agreement. Similarly, in a thorough review of five leading animal behaviour journals during the last five decades, Burghardt et al. (2012) found that <10% of empirical research reported methods for reducing observer bias, and only 3.5% reported any form of statistical reliability measures. Clearly, while most would agree that reliable measurements are a critical component of any scientific research, it appears that much research in animal behaviour generally seems to ignore this issue.

Reliability as a concept includes several related aspects, including "blind" measurements, test-retest consistency, and internal consistency of aggregate measurements (Burghardt et al., 2012; Caro et al., 1979). Here, we focus on two major components of reliability, inter-observer agreement and criterion validity (John & Soto, 2007). Inter-observer agreement (hereafter, "agreement") measures the concordance between codings or ratings of behaviour of the same subject made by different observers. Strong agreement between observers suggests that measures are indeed characteristics of the subject animal, rather than characteristics related to the observer (Gosling, Kwan, & John, 2003; Kaufman & Rosenthal, 2009). Criterion validity (hereafter, "validity") is an index of how accurate a measurement is in describing what it is aiming to measure (Vazire, Gosling, Dickey, & Schaprio, 2007).

Reliability of measures are more commonly reported in studies on human psychology and medical research (Gosling, 2001). Perhaps not surprisingly, within the human psychological literature, there is a framework on how to study measurement reliability (see also Caro et al., 1979 for a less well-defined, but synonymous framework for ethology). Funder (1995) proposed four broad categories where differences in agreement between observers can arise: *good judge*, that some people might be better observers than others; *good target*, the possibility that some subjects are easier to observe than others; *good trait*, that some traits (or behaviours) are easier to observe (or predict) than others; and *good information*, that certain kinds of information (or definitions) make observation more accurate. The issue of "good judge" (i.e., the impact of observers' experience) is of particular concern to a number of behavioural studies (Duncan & Pillay, 2012; Phythian, Michalopoulou, Duncan, & Wemelsfelder, 2013; Tami & Gallagher, 2009). For example, researchers are often cautious of using observers unfamiliar with the study species (Meagher, 2009; Petelle & Blumstein, 2014). However, some research suggests that while reliability of measures increases with level of observers' experience with the study animal, observers less acquainted with subjects can also obtain satisfactory agreement amongst themselves and others considered more "expert" (Fratkin et al., 2015; Martau, Caine, & Candland, 1985; Wemelsfelder, Hunter, Mendl, & Lawrence, 2000).

In this study, we investigated the reliability of ratings used to measure the behaviour of odour-detection working dogs (*Canis familiaris*) in standardized tests using Funder's (1995) framework. Idiosyncratic variation in dog behaviour is well-established (Jones & Gosling, 2005; Sinn, Gosling, & Hilliard, 2010; Svartberg, Tapper, Temrin, Radesater, & Thorman, 2005), and characterization of dog personality is of great theoretical (e.g., Bensky, Gosling, & Sinn, 2013) and applied concern (e.g., Sinn et al., 2010). Dog personality is commonly measured using ratings, which typically consist of assigning a Likert-scale value (such as 1–5) for a behavioural trait based on observed behaviours, for example, "confidence" or "aggressiveness" (Sinn et al., 2010; Wilsson & Sinn, 2012). While some studies have demonstrated that ratings can have high agreement, the same studies also report that some ratings do not (Gosling, 1998, 2001; Gosling et al., 2003; King, Weiss, & Sisco, 2008; Ley, McGreevy, & Bennett, 2009; Sinn et al., 2010; Uher & Asendorpf, 2008; Wielebnowski, 1999). Despite the observation of variable agreement for different ratings (within and across studies), little is known concerning any broad trends. Here, we apply Funder's (1995) framework in an attempt to systematically identify areas of observer bias in animal behavioural research. We tested: (a) whether inter-observer agreement was strongly influenced by the ratings themselves (i.e., Funder's "good trait"), (b) whether the age of the dog influenced agreement of the rating (i.e., "good target"), and (c) whether minimally trained novice ratings matched those given by experts (i.e., our measure of validity, or in Funder's terminology "good judge").

2 | METHODS

2.1 | Subjects

A total of 38 Labrador retrievers from the Australian Custom Service and Border Patrol Detector Dog program (hereafter "AUS Custom") were given standardized behavioural assessments at 3, 6, 9 and 12 months of age from November 2012 to January 2013. These assessments are used to guide decisions regarding the fate of dogs, either as AUS Custom detector dogs or pets. Sample sizes for each of the four ages varied due to the working dog program's logistical constraints, so from the pool of 38 dogs, 10 were assessed at

3 months, 6 at 6 months, 13 at 9 months, and 9 at 12 months. No dog contributed data to more than one age.

Human participants included an independent expert dog handler (“expert”) and six university undergraduate students (“novices”). The independent “expert” (ST) had over 20 years of professional experience working with and assessing dogs, was very familiar with the AUS Custom's behavioural assessment prior to the study and had used a similar assessment tool in a professional setting previously for over a decade. The novice group had no professional experience working with dogs and no previous training with the scientific measurement of dog behaviour beyond the training provided as part of this study (see below). All novices had previous experience owning dogs as pets (mean ownership = 17 years; *SD* = 9.49).

2.2 | Standardized behavioural assessments and ratings

The behaviours observed and rated in the assessments differed across ages. Dogs aged 3 and 6 months were given seven ratings, whereas 9- and 12-month-old dogs were given five ratings. All ratings were given on a 5-point Likert scale. Only full points were given at 3 and 6 months; half points were used at 9 and 12 months. A lower rating indicated a lack of a particular behaviour, whereas a higher rating indicated that a dog performed a particular behaviour more so during an assessment. The seven ratings at 3 and 6 months described the dog's ability to chase and retrieve a scented towel (“chase retrieve”), its ability to interact and hold a scented towel (hereafter “towel”) on its own (“independent possession”), its desire to carry and grab onto a towel during tug-of-war (“physical possession”), its ability to find a hidden towel (“hunt grass”, “hunt 1”, and “hunt 2”) as well as its overall activity level (“activity”). The five ratings given at 9 and 12 months described the dog's eagerness to find a hidden towel (“eagerness to retrieve”), its reliance on olfactory cues when searching a line of pots for a hidden towel (“investigative search”), the level of assistance the dogs needed from the handler to locate a hidden towel (“independence”), the dog's determination and persistence when searching for a towel (“hunt drive”), and the dog's level of enthusiasm and time it took for the dog to pick up a towel once it had been located (“recovery of aid”). See supporting information (Table S1) for more detailed description of the behavioural assessments. The differences in the two measurement instruments for younger (3 and 6 months) versus older dogs (9 and 12 months) reflected AUS Custom's desire to subject older dogs to a semi-realistic context that could be used during the final stages of their selection process.

Different AUS Custom dog trainers handled the dogs during the behavioural assessments and the assessments lasted for approximately 10–15 min and occurred in the same environment at each age. The behavioural assessments were videotaped using a camera (GoPro HD HERO2) attached to the AUS Custom dog trainer's head. The novices and our independent expert rated the dogs at a later date after watching video recordings.

2.3 | Training of novices and the independent expert

Novices and the independent expert (hereafter, “human participants”) were individually given brief training prior to contributing data for analysis. Based on our own experiences, we believe the level of training given to novices here was similar to what is often provided to new participants and volunteers used in animal behavioural research. A standardized form with the description of each of the behaviours assessed at the different ages was provided to each of the participants together with a verbal explanation by the first author (KLM) for each of the behaviours (see Table S1). Human participants were then shown eight training videos that consisted of two dogs at each age and were told what ratings the AUS Custom's dog trainers had assigned the dogs for each of the behaviours. During the training, participants were encouraged to ask questions, rewind, pause and re-watch the videos as much as they needed. Next, each participant was given 12 additional training videos, three videos of dogs at each age, and were asked to assign ratings to the dogs in the video. Videos were watched in age intervals with 3-month-old dogs shown first, then 6-, 9-, and 12-month-old dogs. After the additional training videos, human participants were informed what ratings the AUS Custom dog trainers had assigned to videos, and their scores were compared. We set a minimum match criterion of 70% absolute agreement between AUS Custom's dog trainers’ ratings and our participants’ ratings; all participants met this criterion after the described training, which lasted approximately 2 hr. For data collection, novices and the independent expert watched the videos individually in assigned random order using a standardized form. For all videos, participants were allowed to watch the videos at their own pace, and rewind, pause and re-watch the videos, but were instructed not to talk about their ratings with other participants during the study period. Participants completed their ratings in approximately 1 month.

2.4 | Ethical note

All procedures in this study were in accordance with the human ethics committee at the University of Tasmania (H0017115) and with the 1964 Helsinki declaration. Subject dogs were completely under the care of government veterinarians (AUS Custom detector dog program), and research did not include authors’ interactions with or manipulations of dog subjects.

2.5 | Statistical analysis

We used the intra-class correlation coefficient (ICC [2, *k*]; Shrout & Fleiss, 1979) to test for agreement between novices for each rating at each age and considered ICC coefficients ≥0.70 as indicating “acceptable” agreement, and those <0.70 as indicating “poorer” agreement (Cicchetti, 1994; John & Soto, 2007). There were several cases where there was no variation in a rating assigned to the same dog at the same age, both within and between two or more novices. Due to the lack of variation in a rating (indicating either high agreement or low variation in dog behaviour or both), ICCs could not be estimated. There were three of 48 estimates where this occurred and, in these cases, we used percentage agreement to index agreement. However, as the number of ratings with insufficient variation was small compared to the overall number of ratings, and because percentage agreement does not correct for agreement that would be expected by chance alone (Bartko, 1991; Hallgren, 2012), we report these estimates but do not discuss them further.

Validity estimates of novice ratings compared to our independent expert ratings were evaluated using Pearson's correlations. Validity estimates for each of the ratings given at each age were calculated by first correlating the expert's ratings with each individual novice rating, then averaging the observed single correlation coefficients using Fisher's *r* to *z* formula (Snedecor & Cochran, 1980).

We tested whether agreement between novices was different between ratings (i.e., Funder's "good trait") by averaging agreement coefficients for each rating across all ages. For each rating at each age, we converted ICC estimates to a *z*-value, averaged them across ages for that rating, then reverse-transformed the average *z* to an average *r* with associated 95% confidence intervals (Fisher, 1915). Statistically, differences between ratings' average agreement were assessed against the criterion of overlapping confidence intervals (CIs). When $n \geq 10$, if CI error bars overlap by half, $p \approx 0.05$. If the tips of the error bars just touch, $p \approx 0.01$ (Cumming, Fidler, & Vaux, 2007). We considered any CI error bar overlap >50% to indicate $p > 0.05$.

Similarly, to test whether the age of the dog impacted on the agreement of observers' ratings (i.e., Funder's "good target"), we averaged novice agreement estimates from all ratings within a single age using Fischer's *r*-to-*z* and compared average agreement for ratings amongst different ages by observing overlap of 95% CIs. As different measurement instruments were used at 3/6 months and 9/12 months, we limited our age comparisons to those using the same instrument (i.e., we compared 3 months to 6 months and 9 months to 12 months only).

Finally, to test whether the experience level of the observer affected specific ratings or ages (i.e., Funder's "good judge"), we used Fisher's *r*-to-*z* to average validity estimates within a rating across ages, and amongst ratings within an age, respectively, and used 95% confidence intervals to assess any differences in validity estimates across ratings and amongst ratings across ages. Statistical analyses were conducted in R version 3.3.0 (R Core Team, 2016) using the irr package (Gamer, Lemon, Fellows, & Singh, 2012).

3 | RESULTS

3.1 | Funder's "good trait"

The grand mean of novice agreement estimates for all ratings at 3 and 6 months of age was high ($r = 0.86$, CI = 0.80–0.90), as was the grand mean for all ratings during 9- and 12-month assessments ($r = 0.86$, CI = 0.80–0.90; Table 1). At 3 and 6 months of age, agreement ranged from 0.53 ("hunt grass") to 0.98 ("hunt 1"). At 9 and 12 months of age, agreement ranged from 0.64 ("recovery of aid") to 0.96 ("investigative search"). Within this overall high agreement amongst novices, there were, however, statistically significant differences in average agreement between some of the ratings. Across 3- and 6-month ratings, three ratings with the lowest observed agreement ("physical possession", "independent possession" and "hunt grass") had complete non-overlap of CIs with "hunt 2",

TABLE 1 Agreement amongst six novice observers for behavioural ratings of working dogs

Behaviour	3 month (95% CI)	6 month (95% CI)	Rating average (95% CI)
Chase retrieve	90%	83%	–
Physical possession	0.73 (0.51 to 0.91)	0.82 (0.44 to 0.97)	0.78 (0.67 to 0.85)
Independent possession	0.86 (0.65 to 0.96)	0.88 (0.64 to 0.98)	0.87 (0.85 to 0.89)
Hunt grass	0.53 (−0.15 to 0.86)	0.80 (0.37 to 0.97)	0.69 (0.33 to 0.87)
Hunt 1	0.98 (0.95 to 0.99)	0.71 (0.11 to 0.96)	0.92 (0.21 to 0.99)
Hunt 2	0.94 (0.87 to 0.98)	0.96 (0.88 to 0.99)	0.95 (0.93 to 0.97)
Activity	70%	0.60 (−0.25 to 0.94)	–
Age average	0.88 (0.65 to 0.96)	0.83 (0.69 to 0.91)	Grand $\bar{x}$ = 0.86 (0.80 to 0.90)
Behaviour	**9 month (95% CI)**	**12 month (95% CI)**	**Rating average (95% CI)**
Eagerness to retrieve	0.86 (0.69 to 0.95)	0.86 (0.63 to 0.96)	0.86 (0.86 to 0.86)
Investigative search	0.96 (0.92 to 0.99)	0.91 (0.76 to 0.98)	0.94 (0.87 to 0.97)
Level of independence	0.80 (0.56 to 0.93)	0.81 (0.52 to 0.95)	0.81 (0.80 to 0.81)
Hunt drive	0.87 (0.71 to 0.95)	0.86 (0.66 to 0.97)	0.87 (0.85 to 0.87)
Recovery of aid	0.81 (0.59 to 0.93)	0.64 (0.09 to 0.91)	0.74 (0.52 to 0.86)
Age average	0.88 (0.79 to 0.93)	0.83 (0.74 to 0.90)	Grand $\bar{x}$ = 0. 86 (0.80 to 0.90)

Note. Estimates of agreement coefficients using the intra-class correlation coefficient (ICCs) are given with their associated 95% confidence intervals (CIs) in parentheses. Percentage agreement coefficients were calculated when two or more observers had zero variability in ratings for a particular item across and within dogs. Percentage agreement estimates do not contribute to average estimates reported.

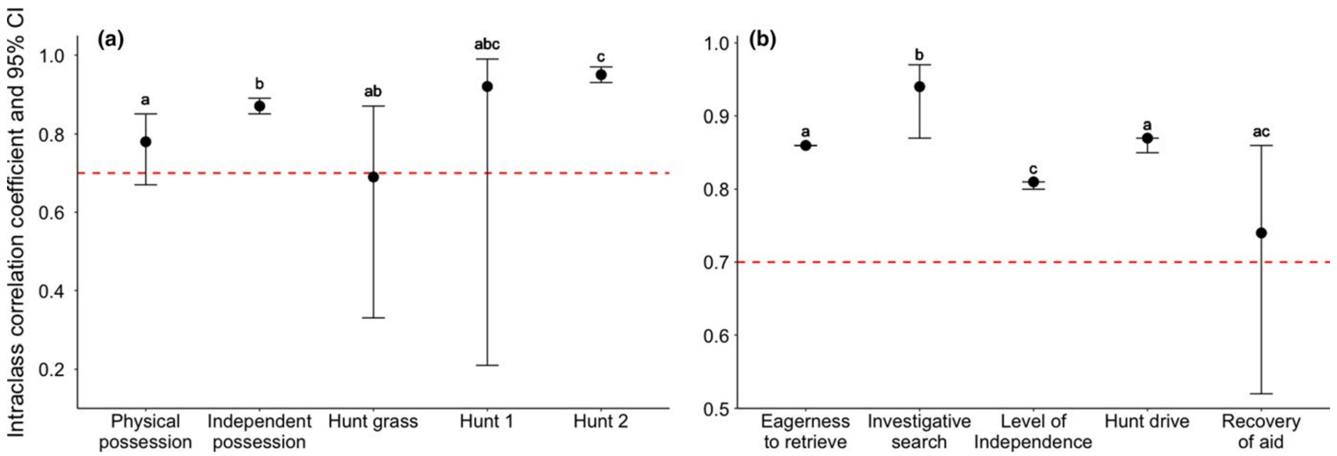

FIGURE 1 Average inter-observer agreement amongst novices (intra-class correlation coefficient and associated 95% confident intervals) for behaviour ratings across (a) 3- and 6-month-old dogs, and (b) 9- and 12-month-old dogs. Different letters indicate that there was no overlap in confidence intervals. Different letters show statistical significance. The dashed line represents "acceptable validity"

TABLE 2 Validity estimates of six novice observers' behavioural ratings of working dogs when compared to an independent expert dog trainer

Behaviour	3 month (95% CI)	6 month (95% CI)	Rating average (95% CI)
Chase retrieve	0.88 (0.58 to 0.97)	0.88 (0.23 to 0.99)	0.88 (0.88 to 0.88)
Physical possession	0.84 (0.43 to 0.96)	0.50 (−0.53 to 0.93)	0.71 (0.22 to 0.92)
Independent possession	0.91 (0.65 to 0.98)	0.53 (−0.49 to 0.94)	0.79 (0.14 to 0.96)
Hunt grass	0.31 (−0.39 to 0.78)	0.29 (−0.68 to 0.89)	0.30 (0.28 to 0.32)
Hunt 1	0.67 (0.06 to 0.91)	0.87 (0.19 to 0.99)	0.79 (0.51 to 0.92)
Hunt 2	0.91 (0.67 to 0.98)	0.99 (0.98 to 0.99)	0.97 (0.76 to 0.99)
Activity	0.41 (0.72 to 0.93)	0.93 (0.72 to 0.98)	0.78 (−0.15 to 0.98)
Age average	0.77 (0.57 to 0.89)	0.84 (0.51 to 0.95)	Grand x̄ = 0.81 (0.73 to 0.87)
Behaviour	**9 month (95% CI)**	**12 month (95% CI)**	**Rating average (95% CI)**
Eagerness to retrieve	0.28 (−0.32 to 0.72)	0.16 (−0.56 to 0.75)	0.22 (0.10 to 0.33)
Investigative search	0.73 (0.29 to 0.91)	0.66 (−0.01 to 0.92)	0.70 (0.62 to 0.76)
Level of independence	0.09 (−0.48 to 0.61)	0.71 (0.10 to 0.94)	0.45 (−0.28 to 0.85)
Hunt drive	0.50 (−0.06 to 0.83)	0.67 (0.02 to 0.92)	0.59 (0.40 to 0.73)
Recovery of aid	0.38 (−0.22 to 0.77)	0.22 (−0.52 to 0.77)	0.30 (0.14 to 0.99)
Age average	0.42 (0.17 to 0.62)	0.52 (0.26 to 0.71)	Grand x̄ = 0.47 (0.37 to 0.56)

Note. Estimates of validity using the Pearson's are given with their associated 95% confidence intervals (CIs) in parentheses.

the rating with the second highest observed agreement ($p < 0.05$; Figure 1a). Furthermore, there was non-overlap of CIs between the two ratings assessing a dog's ability to possess a scented towel ("physical possession" and "independent possession") ($p < 0.05$; Figure 1a). Across 9- and 12-month ratings, the ratings with the lowest average inter-observer agreement ("eagerness to retrieve", "level of independence", "hunt drive" and "recovery of aid") had no overlap of CIs with the rating with the highest agreement ("investigative search") ($p < 0.05$; Figure 1b). Furthermore, there was non-overlap of CIs between the rating assessing a dog's "level of independence" and the two ratings related to a dog's "eagerness to retrieve" and "hunt drive" ($p < 0.05$; Figure 1b).

3.2 | Funder's "good target"

There was no evidence that the age of the dog influenced observer agreement. The average novice agreement across ratings within each age met or exceeded 0.80. Agreement estimates at 3 and 6 months of age had completely overlapping CIs. This was similar for agreement estimates at 9 and 12 months of age (Table 1).

3.3 | Funder's "good judge"

The grand mean of validity estimates across ratings at 3 and 6 months was high ($r = 0.81$; CI = 0.73–0.87), while the grand

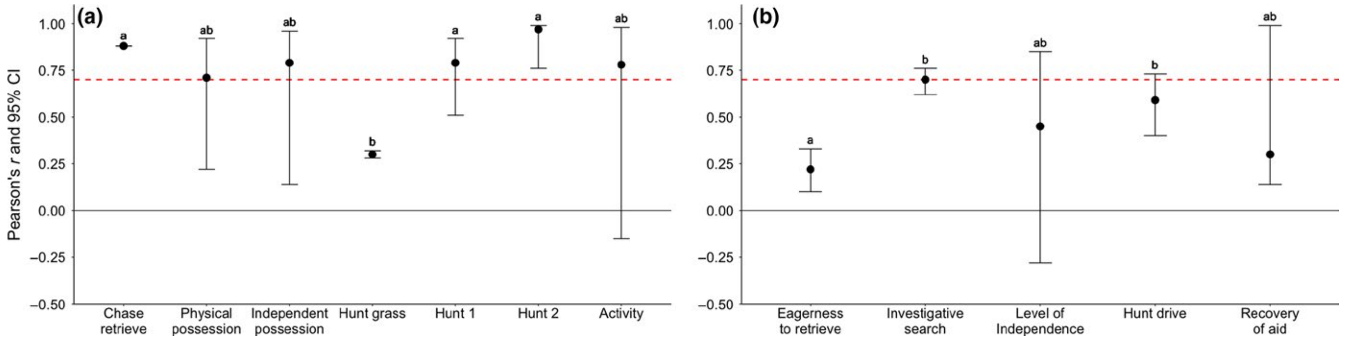

FIGURE 2 Validity estimates (Pearson's *r* and 95% confidence intervals) for behaviours assessed across (a) 3- and 6-month-old dogs, and (b) 9- and 12-month-old dogs, when novice ratings were compared to those given by experts. Different letters indicate statistical significance. The dashed line represents "acceptable validity"

mean of validity across ratings at 9 and 12 months was much lower (r = 0.47; CI = 0.37–0.56; Table 2). Validity estimates for ratings ranged from 0.29 ("hunt grass") to 0.99 ("hunt 2") at 3 and 6 months and ranged from 0.09 ("level of independence") to 0.73 ("investigative search") at 9 and 12 months. At 3 and 6 months, the rating with the lowest validity, "hunt grass" (r = 0.30) had non-overlapping CI with "chase retrieve" (r = 0.88), "hunt 1" (r = 0.79), "hunt 2" (r = 0.97) and "activity" (r = 0.78) (p < 0.05; Figure 2a). At 9 and 12 months, the validity estimate for "eagerness to retrieve" (r = 0.22) had no overlap in CIs with "investigative search" (r = 0.70) and "hunt drive" (r = 0.59) (p < 0.05; Figure 2b).

The average validity estimates between expert and novice ratings at each age (an interaction between "good judge" and "good trait") ranged from strong to poor. Validity was strong at 3 months (0.77) and 6 months (0.84), but poor at 9 (0.42) and 12 months (0.52; Figure 3). Validity estimates for ratings across ages such as "hunt 2" and "activity" increased significantly from 3 to 6 months, whereas validity for "recovery of aid" significantly decreased from 9 to 12 months (p < 0.05; Table 2). In general, validity estimates from ratings by the expert and novices for 9- and 12-month-old dogs were poor compared to those for 3 and 6-month-old dogs. Four of seven ratings (57%) at 3 months had acceptable validity, and four of seven ratings (57%) at 6 months, but only one of five ratings (20%) given at 9 months as well as 12 months, showed acceptable agreement between observers of different experience levels (Table 2).

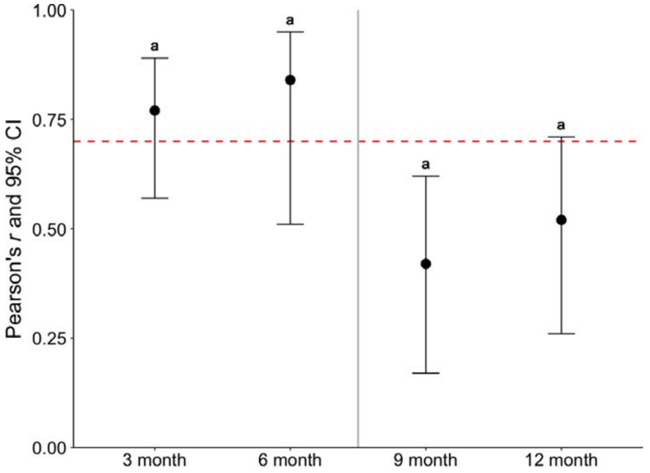

FIGURE 3 Average validity estimates (Pearson's *r* and 95% confident intervals) for behaviours assessed in 3-, 6-, 9- and 12-month-old dogs when novice ratings were compared to those given by experts. The dashed line represents "acceptable validity", whereas the vertical grey line delineates between the two different measurement instruments

4 | DISCUSSION

In 1995, David Funder described four phenomena that can affect agreement between observers in behavioural studies: (a) *good trait*, the possibility that some traits (or behaviours) might be more easily judged than others, (b) *good target*, the possibility that some subjects might be more easily judged than others, (c) *good judge*, the possibility that some observers are better judges than others, and (d) *good information*, the possibility that more or certain kinds of information makes judging more accurate. For the first time, we extend this framework, intended for human psychology, to assess the ways in which reliability can arise in animal behavioural research. We found that observers with the same level of experience can reliably assess a wide range of different types of dog behaviours, i.e., there was no strong evidence of Funder's "good trait". Although in some cases, agreement coefficients were different (i.e., "hunt grass" and "activity" at 3/6 months; "eagerness to retrieve" at 9/12 months), the majority of behaviours (86%; 19 of 22) measured at all ages had acceptable (>0.70), and in many cases, very high inter-observer reliability. We found no strong evidence for Funder's "good target"; our result showed that inexperienced observers tended to agree in their ratings when judging dogs at different ages and using different measurement instruments. We did, however, find effects of a "good judge". The validity of novice ratings when compared to ratings given by an independent expert appeared to be test specific with acceptable validity (>0.70) observed only for 3- and 6-month-old dogs, whereas validity was especially poor for dogs at 9 and 12 months of age using a different measurement instrument.

Novice observers agreed strongly on their ratings for all behaviours. This result fits with the general pattern found in some studies (e.g., Ley et al., 2009), but not others (e.g., Dutton, 2008; Fratkin et al., 2015) which report a mix of estimate effect sizes, some acceptable and others lacking in reliability. For example, Dutton (2008) reported high inter-observer agreement for traits such as dominance, playfulness and neuroticism, whereas deceptiveness, trust and social awareness, arguably subtler in nature, had low agreement in a study of captive chimpanzees (*Pan troglodytes*). Fratkin et al. (2015) suggested that good agreement amongst observers may be more difficult to attain for traits that involve a high degree of visual, physical, or tactile subtlety (see also Kristensen et al., 2006). In other words, the differences in agreement amongst observers between traits/behaviours may arise from the available visual cues that are easily interpreted by human observers (Funder, Kolar, & Blackman, 1995; Gosling, 2001; John & Srivastava, 1999). Further empirical studies, along with subsequent meta-analyses, are needed to identify any general patterns in how or why some behaviours are easier to observe than others, and whether this is taxa specific (see Gosling, 2001).

Assessing inter-observer agreement across different developmental ages of subject animals is rare in animal behaviour (Kaufman & Rosenthal, 2009) including in dogs (but see Sinn et al., 2010; Fratkin et al., 2015). The high agreement between the novices in our study at both 3/6 and 9/12 months may be partly explained because our subject animal was the domesticated dog. Human–dog relationships began almost 15,000 years ago (Bokkers, de Vries, Antonissen, & de Boer, 2012; Kaminski, Braeuer, Call, & Tomasello, 2009; Riedel, Schumann, Kaminski, Call, & Tomasello, 2008), resulting in both species becoming mutually perceptive of one another. High inter-observer agreement amongst novice observers has also been observed in other domesticated animals, such as cows and pigs (Meagher, 2009; Wemelsfelder, Hunter, Paul, & Lawrence, 2012). While it is reassuring to know that oftentimes, at least with domesticated animals, novice observers of animal behaviour tend to agree, it is also true that due to the current lack of data from non-domesticated species, it is unknown how novices might (dis)agree on the behaviour that they observe in other taxa. Studies that examine agreement between observers in non-domesticated species are needed, which would allow for exciting future research questions regarding the differences in human perception based on subject animals' evolutionary background.

When comparing novice ratings to those given by our independent expert (our measure of validity), the age of the dog did not appear to influence the agreement between the two types of observers when limiting the comparison to only 3 versus 6 months and 9 versus 12 months ratings. Ratings given to both 3- and 6-month-old dogs had acceptable validity (validity coefficient = 0.77 and 0.84 respectively), whereas ratings given to 9-month-old and 12-month-old dogs were low (0.42 and 0.52, respectively). Given that the age of the dogs and the measurement instrument used were confounded, we are unable to differentiate whether this result is caused by an age effect, differences in the measurement instrument or a combination of both. There is some evidence to support the idea that the differences in validity were determined by the change in the measurement instrument at 9 and 12 months. In a comparable study, Fratkin et al. (2015) utilized the same measurement instrument (similar to the one used here at 3 and 6 months) across 3-, 6-, 9-, and 12-month-old dogs and reported consistent but moderate (>0.54 but <0.66) validity for ratings across all ages. On the other hand, Funder (1995) argued that the differences between observer accuracy are produced by differential detection or weight that different observers give to available cues. In our case, the working dog expert, with decades of training and experience, may have been able to perceive behaviours in the dogs that the novices were unable to do (e.g., postural and other subtle cues; Kristensen et al., 2006). Given the goal of many university research programs and animal behaviour laboratories to involve and utilize novices (e.g., undergraduates: Eagan, Sharkness, Hurtado, Mosqueda, & Chang, 2011), it is worth noting that our study along with others suggests that while novices may agree with themselves, they do not necessarily agree with an expert (Hróbjartsson et al., 2013). Further research on the validity of behavioural measures given by different groups of people with different levels of experience is clearly needed.

While our study design did not allow for a clear test Funder's fourth category, "good information", it is worth noting two general themes in current animal behaviour methods that are relevant in this respect. First, there are generally two different methods used to record animal behaviour: behavioural codings and behavioural ratings (Gosling, 2001; Martin & Bateson, 1993; Wilsson & Sinn, 2012). Behavioural coding methods attempt to narrowly define discrete, observable behaviours and then use frequency counts and durations of those observed behaviours to code subject behaviour; behaviour codings are often based on species or population ethograms. Ratings methods, on the other hand, rely on human observers to intuitively aggregate and interpret behaviour for particular, pre-defined traits. Ratings are usually based on Likert scales, for example, from one to five or one to seven, where the number reflects the severity of the behaviour displayed by the subject animal as deemed by a human observer. Ratings can be further delineated by the level of detail given to the observer in terms of what each number in a scale represents. For example, some ratings give rigid detail for each level of a scale (e.g., Godfrey, Bradley, Sih, & Bull, 2012), while other ratings, such as ours used here, require the observer to use more intuition as to what "more" or "less" of a behaviour means (e.g., Wilsson & Sinn, 2012). Having multiple observers give behavioural codings and ratings to the same individuals or giving different ratings (some with more detailed information than others) could provide tests as to how different measures with different types of embedded information could yield higher/lower reliability or validity. Second, video methods of animal behaviour studies are common. For some types of behavioural measures (such as those observed here), video methods may not matter. However, in some cases, inter-observer agreement or criterion validity may also depend on whether observers are allowed to watch animals in "real-time" or whether they measure behaviour from video at a later date. For videos, subtle behaviours or physical signals might be diluted or completely lost (Diesel, Brodbelt, & Pfeiffer, 2008; Wells & Hepper, 1999).

In general, tests of "good information" are also in need in animal behavioural studies.

In the majority of animal behaviour studies, researchers have failed to demonstrate that different observers watching the same animal agree on what the animal is doing. We attempted to fill this gap by adopting a framework intended for human psychology (Funder, 1995) and applied it to animal research. This allowed us to move beyond simply reporting reliability and instead begin to address the ways in which reliability (or lack of) may arise in animal behavioural research. Interestingly, in our study, novices strongly agreed with one another, but novices were not always in agreement with the independent expert. It is currently unknown how widespread this phenomenon may be, but further tests of how different groups of people may (dis)agree about behaviour across different taxa are needed. This is especially pertinent as animal behaviour programs attempt to attract undergraduate volunteers and members of the public, and to long-term research projects that utilize a range of observers with different levels of expertise.

ACKNOWLEDGEMENTS

We thank Australian Custom and Border Protection for providing video recording of their behavioural assessments as well as their ratings used for training purposes, and the six novice volunteers for providing ratings. We are thankful to Glen Scott and Smyl Fisher for support and access to videos and ratings of Australian Border Force Detector dogs. DS was supported by a grant from US Department of Homeland Security - Science and Technology Directorate (HSHQDC-10-C-00085), and EW acknowledges the support of the Australian Research Council.

CONFLICT OF INTEREST

The authors declare no competing interests.

ORCID

Kirke L. Munch https://orcid.org/0000-0003-2929-6805

REFERENCES

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin*, *17*, 483–489. https://doi.org/10.1093/schbul/17.3.483

Bensky, M. K., Gosling, S. D., & Sinn, D. L. (2013). The world from a dog's point of view: A review and synthesis of dog cognition research. *Advances in the Study of Behavior*, *45*, 209–406.

Bokkers, E. A. M., de Vries, M., Antonissen, I. C. M. A., & de Boer, I. J. M. (2012). Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. *Animal Welfare*, *21*, 307–318. https://doi.org/10.7120/09627286.21.3.307

Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrison, K. E., Stec, C. L., Zachau, C. E., & Freeberg, T. M. (2012). Perspectives—minimizing observer bias in behavioral studies: A review and recommendations. *Ethology*, *118*(6), 511–517. https://doi.org/10.1111/j.1439-0310.2012.02040.x

Caro, T. M., Roper, R., Young, M., & Dank, G. R. (1979). Inter-observer reliability. *Behaviour*, *69*(3/4), 303–315.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, *177*(1), 7–11. https://doi.org/10.1083/jcb.200611141

Diesel, G., Brodbelt, D., & Pfeiffer, D. U. (2008). Reliability assessment of dogs' behavioural responses by staff working at a welfare charity in the UK. *Applied Animal Behaviour Science*, *115*, 171–181.

Duncan, L. M., & Pillay, N. (2012). Volunteer experience influences the conclusions of behavioural experiments. *Applied Animal Behaviour Science*, *140*(3–4), 179–187. https://doi.org/10.1016/j.applanim.2012.06.003

Dutton, D. M. (2008). Subjective assessment of chimpanzee (*Pan troglodytes*) personality: Reliability and stability of trait ratings. *Primates*, *49*, 253–259. https://doi.org/10.1007/s10329-008-0094-1

Eagan, M. K., Sharkness, J., Hurtado, S., Mosqueda, C. M., & Chang, M. J. (2011). Engaging undergraduates in science research: Not just about faculty willingness. *Research in Higher Education*, *52*(2), 151–177. https://doi.org/10.1007/s11162-010-9189-9

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*, 507–521.

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? *Behavioural Processes*, *110*, 105–116.

Funder, D. C. (1995). On the accuracy of personality judgement: A realistic approach. *Psychological Review*, *102*, 652–670.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintance. *Journal of Personality and Social Psychology*, *69*, 656–672.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Irr: Various coefficients of interrater reliability and agreement. [computer software]*. Retrieved from https://cran.r-project.org/web/packages/irr/index.html

Godfrey, S. S., Bradley, J. K., Sih, A., & Bull, C. M. (2012). Lovers and fighters in sleepy lizard land: Where do aggressive males fit in a social network? *Animal Behaviour*, *83*, 209–215.

Gosling, S. D. (1998). Personality dimensions in spotted hyenas (*Crocuta crocuta*). *Journal of Comparative and Physiological Psychology*, *112*, 107–118. https://doi.org/10.1037/0735-7036.112.2.107

Gosling, S. D. (2001). From mice to men: What can we learn about personality from animal research? *Psychological Bulletin*, *127*(1), 45–86. https://doi.org/10.1037//0033-2909.127.1.45

Gosling, S. D., Kwan, V. S. Y., & John, O. P. (2003). Dog's got personality: A cross-species comparative approach to personality judgments in dogs and humans. *Journal of Personality and Social Psychology*, *85*(6), 1161–1169. https://doi.org/10.1037/0022-3514.85.6.1161

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., ... Brorson, S. (2013). Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *Canadian Medical Association Journal*, *185*, E201–211. https://doi.org/10.1503/cmaj.120744

John, O. P., & Soto, C. J. (2007). The importance of being valid. In R. Robins, R. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York, NY: The Guilford Press.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research*. New York, NY: Guilford Press.

Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hammerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change* (pp. 7–67). Champaigne, IL: Research Press.

Jones, A. C., & Gosling, S. D. (2005). Temperament and personality in dogs (*Canis familiaris*): A review and evaluation of past research. *Applied Animal Behaviour Science*, *95*(1–2), 1–53. https://doi.org/10.1016/j.applanim.2005.04.008

Kaminski, J., Braeuer, J., Call, J., & Tomasello, M. (2009). Domestic dogs are sensitive to a human's perspective. *Behaviour*, *146*(7), 979–998. https://doi.org/10.1163/156853908x395530

Kaufman, A. B., & Rosenthal, R. (2009). Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. *Animal Behaviour*, *78*(6), 1487–1491. https://doi.org/10.1016/j.anbehav.2009.09.014

King, J. E., Weiss, A., & Sisco, M. N. (2008). Aping humans: Age and sex effects in chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*) personality. *Journal of Comparative and Physiological Psychology*, *122*, 418–427. https://doi.org/10.1037/a0013125

Kristensen, E., Dueholm, L., Vink, D., Anderson, J. E., Jakobsen, E. B., Illum-Nielsen, S., … Enevoldsen, C. (2006). Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *Journal of Dairy Science*, *89*, 3721–3728. https://doi.org/10.3168/jds.S0022-0302(06)72413-4

Levitis, D. A., Lidicker, W. Z., & Freund, G. (2009). Behavioural biologists do not agree on what constitutes behaviour. *Animal Behaviour*, *78*(1), 103–110. https://doi.org/10.1016/j.anbehav.2009.03.018

Ley, J. M., McGreevy, P., & Bennett, P. C. (2009). Inter-rater and test–retest reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R). *Applied Animal Behaviour Science*, *119*(1–2), 85–90. https://doi.org/10.1016/j.applanim.2009.02.027

Martau, P., Caine, N. G., & Candland, D. (1985). Reliability of the emotions profile index, primate form, with *Papio hamadryas*, *Macaca fuscata*, and two *Saimiri* species. *Primates*, *26*, 501–505. https://doi.org/10.1007/BF02382466

Martin, P., & Bateson, P. (1993). *Measuring behavior: An introductory guide* (2nd ed.). Cambridge: Cambridge University Press.

Meagher, R. K. (2009). Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science*, *119*(1–2), 1–14. https://doi.org/10.1016/j.applanim.2009.02.026

Petelle, M. B., & Blumstein, D. T. (2014). A critical evaluation of subjective ratings: Unacquainted observers can reliably assess personality. *Current Zoology*, *60*, 162–169.

Phythian, C., Michalopoulou, E., Duncan, J., & Wemelsfelder, F. (2013). Inter-observer reliability of qualitative behavioural assessments of sheep. *Applied Animal Behaviour Science*, *144*(1–2), 73–79. https://doi.org/10.1016/j.applanim.2012.11.011

R Core Team (2016). *R: Language and environment for statistical computering*. Vienna, Austria: R Foundation for Statistical Computering. Retrieved from http://R-project.org/.

Riedel, J., Schumann, K., Kaminski, J., Call, J., & Tomasello, M. (2008). The early ontogeny of human–dog communication. *Animal Behaviour*, *75*(3), 1003–1014. https://doi.org/10.1016/j.anbehav.2007.08.010

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Sinn, D. L., Gosling, S. D., & Hilliard, S. (2010). Personality and performance in military working dogs: Reliability and predictive validity of behavioral tests. *Applied Animal Behaviour Science*, *127*(1–2), 51–65. https://doi.org/10.1016/j.applanim.2010.08.007

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IA: Iowa State University Press.

Svartberg, K., Tapper, I., Temrin, H., Radesater, T., & Thorman, S. (2005). Consistency of personality traits in dogs. *Animal Behaviour*, *69*, 283–291. https://doi.org/10.1016/j.anbehav.2004.04.011

Tami, G., & Gallagher, A. (2009). Description of the behaviour of domestic dog (*Canis familiaris*) by experienced and inexperienced people. *Applied Animal Behaviour Science*, *120*(3–4), 159–169. https://doi.org/10.1016/j.applanim.2009.06.009

Uher, J., & Asendorpf, J. B. (2008). Personality assessment in the great apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality*, *42*(4), 821–838. https://doi.org/10.1016/j.jrp.2007.10.004

Vazire, S., Gosling, S. D., Dickey, A. S., & Schaprio, S. J. (2007). Measuring personality in nonhuman animals. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 190–206). New York, NY: Guilford.

Wells, D. L., & Hepper, P. G. (1999). Male and female dogs respond differently to men and women. *Applied Animal Behaviour Science*, *61*(4), 341–349. https://doi.org/10.1016/s0168-1591(98)00202-0

Wemelsfelder, F., Hunter, A. E., Paul, E. S., & Lawrence, A. B. (2012). Assessing pig body language: Agreement and consistency between pig farmers, veterinarians, and animal activists. *Journal of Animal Science*, *90*, 3652–3665.

Wemelsfelder, F., Hunter, E. A., Mendl, M. T., & Lawrence, A. B. (2000). The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science*, *67*(3), 193–215. https://doi.org/10.1016/s0168-1591(99)00093-3

Wielebnowski, N. C. (1999). Behavioral differences as predictors of breeding status in captive cheetahs. *Zoo Biology*, *18*, 335–349. https://doi.org/10.1002/(SICI)1098-2361(1999)18:4<335:AID-ZOO8>3.0.CO;2-X

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. *Applied Animal Behaviour Science*, *141*, 158–172. https://doi.org/10.1016/j.applanim.2012.08.012

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Munch KL, Wapstra E, Thomas S, Fisher M, Sinn DL. What are we measuring? Novices agree amongst themselves (but not always with experts) in their assessment of dog behaviour. *Ethology*. 2019;00:1–9. https://doi.org/10.1111/eth.12846