Temperament and problem solving in a population of adolescent guide dogs

Emily E. Bray, Mary D. Sammel, Robert M. Seyfarth, James A. Serpell & Dorothy L. Cheney
Temperament and problem solving in a population of adolescent guide dogs

Emily E. Bray, Mary D. Sammel, Robert M. Seyfarth, James A. Serpell, Dorothy L. Cheney

Abstract It is often assumed that measures of temperament within individuals are more correlated to one another than to measures of problem solving. However, the exact relationship between temperament and problem-solving tasks remains unclear because large-scale studies have typically focused on each independently. To explore this relationship, we tested 119 prospective adolescent guide dogs on a battery of 11 temperament and problem-solving tasks. We then summarized the data using both confirmatory factor analysis and exploratory principal components analysis. Results of confirmatory analysis revealed that a priori separation of tests as measuring either temperament or problem solving led to weak results, poor model fit, some construct validity, and no predictive validity. In contrast, results of exploratory analysis were best summarized by principal components that mixed temperament and problem-solving traits. These components had both construct and predictive validity (i.e., association with success in the guide dog training program). We conclude that there is complex interplay between tasks of “temperament” and “problem solving” and that the study of both together will be more informative than approaches that consider either in isolation.

Keywords Temperament · Cognition · Problem solving · Behavior · Canine · Guide dogs

Introduction

Temperament and cognition, or problem solving, are often regarded as distinct phenomena. Although each is difficult to define, temperament is usually associated with affect, or “the probability of experiencing and expressing the primary emotions and arousal” (Goldsmith et al. 1987, p. 510). It is presumed to be genetically based and present from a young age (Goldsmith et al. 1987). An animal’s temperament is generally considered to affect performance in tasks that are emotionally charged and/or occur in unfamiliar situations. In practice, temperament is often measured by gauging the animal’s reaction to startling events (e.g., King et al. 2003) or willingness to engage with novel environments (e.g., Brown et al. 2005; Greenberg 1984). Cognition, in contrast, is typically associated with “all processes involved in acquiring, storing, and using information from the environment” (Shettleworth 2013, p. 1). Cognition is thought to influence performance in tasks that require behavioral flexibility, innovation, and insight, such as novel tool use (e.g., Taylor et al. 2010), decision making (e.g., Dill 1987), problem solving (e.g.,
Benson-Amram et al. 2016), and inhibitory control (e.g., MacLean et al. 2014).

If temperament and problem-solving skills truly correspond to distinct mental processes, we might expect an individual’s performance on a task purporting to test temperament to be more similar to other tasks ostensibly drawing on this same ability than to cognitive tasks that involve problem solving. Indeed, there is evidence that individual performance on tasks within these domains is associated with one another. Tasks of cognition correlate to other tasks of cognition in studies of dogs, rodents, and male bowerbirds (Arden and Adams 2016; Isden et al. 2013; Matzel and Kolata 2010). Similarly, temperament traits (e.g., aggression, activity, and boldness) often co-occur together in “behavioral syndromes” (Sih et al. 2004). These findings leave open the scenario that temperament and cognition are easily separable constructs, but importantly do not explicitly test for it because they’re only ever looking at half of the picture.

And in fact, several authors challenge such a strict dichotomy. One of the earliest examples originates from Pavlov, who proposed that aspects of temperament, or “types of nervous systems,” were determinants of cognition, as measured by individual differences in the speed of acquisition of classical conditioning (Carere and Locurto 2011; Pavlov 1935; Strelau 1997). More recent work supports the connection between temperament and problem-solving skills (Guillette et al. 2017; Rowe and Healy 2014). For example, several studies of innovative problem solving—a putative cognitive skill (Morand-Ferron et al. 2015; Reader 2003)—suggest that certain attributes associated with temperament (e.g., neophobia, exploration) play a nontrivial role in affecting animals’ success (Auersperg et al. 2011; Benson-Amram and Holekamp 2012; Guido et al. 2017; Overington et al. 2011; Webster and Lefebvre 2001). In other words, animals that are bold and exploratory in their interactions with a problem-solving apparatus increase their opportunities to solve it, whereas animals that are too timid even to approach the apparatus greatly diminish their chance of success. In other examples of the interconnectedness between the two domains, point-following in domesticated foxes is affected by selection for temperament traits (e.g., willingness to approach a human) (Hare et al. 2005), and dogs’ performance on a detour task involving inhibitory control varies based on temperament (Bray et al. 2015).

Large-scale studies purporting to measure problem-solving performance across species have typically not attempted to account for the possible contributions of temperament. For example, several studies have attempted to measure the relationship between brain size and putative cognitive skills in different species of birds (response to novel environments: Sol et al. 2005), carnivores (problem solving: Benson-Amram et al. 2016), and both birds and mammals (inhibitory control: MacLean et al. 2014). None, however, included analyses about the possible contributions of attributes like motivation and neophobia to performance. Similarly, Svarberg and Forkman (2002) conducted a behavioral study of temperament in over 15,000 dogs, but did not collect measures on problem solving within the same animals. While each of these studies is useful in its own right, there has been a lack of large-scale empirical studies in which animals are given multiple behavioral tasks designed to measure both what we might call temperament and cognition (Carere and Locurto 2011; Griffin et al. 2015). It seems apparent both that temperament and problem solving refer to functionally distinct domains, and also that they interact: however, the degree to which temperament and problem solving depend upon one another remains an open question.

For theoretical and practical reasons, dogs are an ideal species in which to study the intersection of these two domains. The utility of dogs in herding, hunting, detecting, serving, and guiding depends on specific behaviors and skills (Hart and Yamamoto 2017; Lord et al. 2017). Many studies have explored cognition in dogs (e.g., Kaminski and Marshall-Pescini 2014; Miklösi 2015) and temperament (e.g., Riemer et al. 2016; Serpell and Hsu 2001)—however, it is rare to see cognition and temperament explicitly examined in the same study (but see Bray et al. 2015; Marshall-Pescini et al. 2008; Passalacqua et al. 2013). The need for such a synthesis is beginning to be acknowledged (Marshall-Pescini et al. 2017). Finally, high levels of community interest facilitate the recruitment of canine participants for behavioral testing (e.g., Riemer et al. 2014) and the collection of survey responses from caretakers (Hsu and Serpell 2003).

Here, we subjected a large sample of prospective adolescent guide dogs to a battery of tasks. Some tasks were designed to elicit responses that we assumed, based on past literature and intuition, would be primarily affected by measures of temperament, such as boldness, reactivity, neophobia, distraction, and fearfulness (e.g., Goddard and Beilharz 1986; Pfaffengerber et al. 1976; Wilsson and Sundgren 1998). Others, we assumed, would elicit responses influenced mainly by problem-solving skills, such as impulse control,1 innovation, and behavioral flexibility (e.g., Barrera et al. 2015; Benson-Amram and Holekamp 2012; Bray et al. 2014). We also measured three

1 Perhaps speaking to the inherent difficulty in separating temperament and cognition, impulse control is sometimes grouped in the ‘temperament’ category within the human literature (e.g., Duckworth and Alfred 2012), although acknowledged to relate to executive control. However, we adhered to the convention in the animal literature, where it is considered a ‘cognitive’ ability (e.g., Amici et al. 2008; Dettmer et al. 2017; MacLean et al. 2014).
dependent variables in order to validate our behaviorally derived factors: subjective assessments of behavior as measured by questionnaires, a physiological measure of stress as measured by salivary cortisol, and success as measured by outcome in the guide dog training program.

We initially performed a confirmatory factor analysis to model the data, permitting us to identify a priori which tasks would load onto two separate factors, named Problem solving and Temperament. When this analysis yielded poor results, we next applied exploratory principal components analysis. This is a data-driven approach in which variables load onto factors based on patterns of correlation within the data without any a priori framework (e.g., Musek 2007). Using these alternative approaches allowed us to ask which mode of analysis best fit our data, and which led to factors that were most valid, as assessed through construct and predictive validity.

We assessed construct validity—i.e., that we were measuring what we intended to—in two different ways. First, we tested whether our behavior-based factors were associated with subjective ratings of similar constructs obtained from puppy-raiser questionnaires. Subjective human ratings have successfully predicted the behavior that they are intended to capture in studies of dogs (Barnard et al. 2016; Foyer et al. 2014; Gosling et al. 2003; Svartberg 2005; although see Brucks et al. 2017b) and other animals (e.g., rhesus macaques: Bolig et al. 1992; baboons: Carter et al. 2012a; gray langurs: Konečná et al. 2008; and horses: Lloyd et al. 2007).

As a second measure of construct validity, we tested whether our behavior-based factors were associated with a physiological measure known to vary across individuals: in this case, cortisol. Cortisol concentration in dogs reflects HPA axis activity and is related to other measures of stress-related behaviors (Beerra et al. 1999; Carrier et al. 2013; Dreschel and Granger 2005; Hydbring-Sandberg et al. 2004). Several past studies have found an association between cortisol levels and various measures of canine temperament (Batt et al. 2009; De Palma et al. 2005; Dreschel and Granger 2005; Hennessy et al. 2001; King et al. 2003; Siniscalchi et al. 2013). In other species, cortisol has been specifically linked to measures of temperament: high levels are associated with excitability in cattle (Burdick et al. 2011; Curley et al. 2006) and fearfulness in primates (Buss et al. 2003; Kalin et al. 1998).

Finally, we tested whether our factors had predictive validity by examining associations between them and later success in the guide dog program. Past studies of working dogs have implicated measures of temperament in successful completion of training (Duffy and Serpell 2012; Harvey et al. 2016; McGarrity et al. 2016; Sinn et al. 2010). One study has also found a measure of problem solving to be predictive (Bray et al. in press).

On the assumption that measures of temperament and problem solving are separable through behavioral testing, we predicted that the confirmatory approach would produce two factors that fit the data well and had strong construct and predictive validity. Specifically, we expected the “temperament” factor to correspond with ratings of excitability, nonsocial fear, and separation-related behaviors, and the “problem-solving” factor to correspond with ratings of trainability and impulsivity. We hypothesized that salivary cortisol concentrations would be associated with the “temperament” factor, and that both factors should be associated with success in the program. Under the same assumption, we predicted that the exploratory approach would reveal similar results, with temperament tasks loading only onto factors with other temperament tasks, and the same with problem-solving tasks.

However, if aspects of temperament and problem solving are more deeply intertwined, then the confirmatory approach should fit poorly, lack associations with previous measures, and fail to be associated with outcome. Additionally, the exploratory approach should reveal factors that do not sort easily into putative temperament and problem-solving categories, but instead reflect a combination of each. Furthermore, these components should correspond to questionnaire and biological (stress-related) measures of temperament and problem solving. They should also be associated with outcome in the program.

General methods

Subjects

Subjects were the same as those described in Bray et al. (in press). Participating dogs were recruited through The Seeing Eye, Inc. (Morristown, NJ, USA), a philanthropic non-profit organization that breeds, raises, and trains guide dogs for the blind and visually impaired. The Seeing Eye granted informed consent to all aspects of the study. All dogs were whelped at the breeding station, weaned at five weeks, and then sent to volunteer puppy raisers at seven weeks. These families were responsible for basic obedience training, taking their puppies to regular meetings with families and dogs in the same geographic region, and exposing their puppies to a wide array of people, animals, locations, and experiences. Between 14 and 17 months, dogs were recalled to headquarters to begin professional training. All testing procedures adhered to regulations set forth by the University of Pennsylvania Institutional Animal Care and Use Committee (Protocol #805210).

We tested 133 young adult dogs from May–October of 2015. However, 14 dogs had to be excluded for missing test scores on at least one task due to video malfunction.
(n = 2) or failure to complete warm-up trials on the memory problem solving, multistep problem solving, and/or cylinder tasks (n = 12). The 26 dogs that were Labrador-Golden crosses were assigned to the breed that contributed over 50% of their genes if applicable, and in all other cases to the breed of their mother. Thus, our final sample consisted of 119 dogs (46 German Shepherds, 55 Labrador Retrievers, and 18 Golden Retrievers) from 21 different litters (Table S1).

**Questionnaires**

Puppy raisers filled out two online questionnaires when the dogs were approximately one year old and had not yet returned to headquarters. They first completed the Canine Behavioral Assessment and Research Questionnaire (C-BARQ®, www.cbarq.org), a behavioral survey designed and validated by Hsu and Serpell (2003) (Appendix A of electronic supplementary material). C-BARQ data were obtained for 114 of the 119 dogs that went on to complete young adult testing (Table S1).

Puppy raisers also completed a 13-item attention deficit hyperactivity disorder (ADHD) rating scale (RS), modified from a human questionnaire (DuPaul et al. 1998) and applied to and validated in dogs (Lit et al. 2010; Vas et al. 2007; Wan et al. 2013) (Appendix B of electronic supplementary material). Puppy raisers read phrases such as “(S)he is quick to break his/her ’rest’ command” and “(S)he follows simple commands easily, such as ’sit’, but (s)he often has difficulties with more complicated commands, such as ’go to your place’, even if (s)he knows them and has practiced them often.” They were then asked to rate the frequency (never, sometimes, often, or very often) with which these statements applied to their dog. Dog-ADHD RS data were obtained for 105 of the 119 dogs that went on to complete young adult testing (Table S1).

**Salivary cortisol collection**

The transition from puppy-raiser homes to a kennel environment is recognized as a particularly stressful time for dogs (Hennessy et al. 1997, 2001; Rooney et al. 2007). We collected saliva after the dogs returned to headquarters and entered the kennels, either one (n = 1), two (n = 102), or three (n = 13) days postarrival and prior to participation in young adult testing. Collection occurred between 7:40 and 8:50 h, at least 1.5 h after the dog’s morning meal. Collection followed previously published methods (Bray et al. 2017; Dreschel and Granger 2009). Briefly, one to two experimenters held the dog while the main experimenter (EB), wearing latex gloves, held a Salimetrics® Children’s Swab under the dog’s tongue and in the dog’s cheek pouches, avoiding contact with the gums, for 1–5 min (Coppola et al. 2006). The swab was gently moved around the dog’s mouth throughout sampling.

Postcollection, all samples (n = 116, Table S1) were stored in a plastic Salimetrics® tube and either immediately refrigerated (at 4 °C) or placed in a freezer (at −20 °C). If samples were refrigerated first, they were frozen no more than 40 min later. One to five months later, samples were mailed on dry ice to Arizona State University’s Institute for Interdisciplinary Salivary Bioscience Research, where they were thawed and assayed for cortisol using ELISA technology and an enzyme immunoassay kit (Salimetrics, Carlsbad, CA). All samples were assayed in duplicate using 25 µL of saliva, and the average of these two measures was used in subsequent analyses. The lower limit of detection was 0.007 µg/dL, and there was an average intra-assay coefficient of variation of less than 10% and an average inter-assay coefficient of variation of less than 15%. To ensure accuracy, analysis was repeated for 18 samples that had a coefficient of variation greater than 15%.

**Young adult testing**

All testing sessions occurred during the initial weeks of return, before each dog was assigned to a professional trainer. Our test battery (described in Bray et al. in press) consisted of 11 tasks designed to capture aspects of dogs’ temperament and problem solving. All dogs were tested on the entire battery twice: they participated in the first testing within the first week of returning to headquarters for training [mean = 2.3 days postarrival, median = 2 days, interquartile range (IQR) = 0 days, range = 1–6 days], and the second testing within a week of undergoing anesthesia for neutering and/or hip X-rays, which occurred over a more variable period after initial return [mean = 12.9 days postarrival, median = 10 days, IQR = 6 days, range = 5–35 days].

On the testing day, tasks were administered over two sessions. Session one lasted approximately one hour, and consisted of the first seven tasks. Dogs were then given at least an hour break before completing session two, which lasted around 30 min and consisted of the final four tasks. For the sake of consistency and intra-individual comparison (MacLean et al. 2017), dogs completed all tasks in the same order.

Dogs (n = 119) were tested at The Seeing Eye® headquarters in an empty 11’ × 7’ tiled exam room located within an unoccupied kennel wing, near the kennel in which the dogs were housed. All tasks took place in the testing room, except for the distraction task, which was conducted in a 44’ × 4’ empty hallway within the same kennel. Testing took place between 7:30 and 17:30 h. For all tasks that required a reward, dogs worked for Zuke’s® mini naturals roasted chicken treats. A main experimenter
and a dog handler were always present. At any given time, these roles were filled by two of five females of similar age, with the first author (EB) acting as the main experimenter in 87% of sessions. Aside from positioning the dog, the handler did not interact with the dog during testing. Testing sessions were videotaped using Sony video cameras (HDR-PJ230, HDR-CX405) mounted on tripods.

Below, we briefly describe each task in the order in which it was presented to the dogs. For further details, see SI methods of electronic supplementary material.

Isolation

This task measured a dog’s comfort level when placed alone in an unfamiliar environment (similar to Gazzano et al. 2008; Wilsson and Sundgren 1998). The handler released the dog into the empty lighted testing room, then left for 2 min. The dependent measures were number of times the dog switched between quadrants of the room over the course of the session and duration of time spent near exit, vocalizing, active, and rearing up on hind legs, respectively. Based on classification in past animal literature (e.g., Kazlauckas et al. 2005; Prut and Belzung 2003; Seaman et al. 2002), we considered this to be primarily a temperament test, measuring activity and anxiety.

Distraction

This task measured a dog’s ability to ignore salient distractors in favor of approaching an encouraging human (Goddard and Beilharz 1984; Harvey et al. 2016). The experimenter called the dog from the other end of a hallway. Initially, the hallway was empty. In subsequent trials, toy and treat distractors were placed along the dog’s route. The dependent measures were time to walk down an empty hallway, number of toys touched, and number of treats eaten. We also calculated a difference score by subtracting each dog’s initial time from her “distraction” time, so that a bigger difference indicated a longer time to complete the “distraction” trials. Based on classification in past animal literature (e.g., Jones and Gosling 2005; Murphy 1998; Weiss and Greenberg 1997), we considered this to be primarily a temperament test, measuring distractibility.

Sustained attention

This task measured a dog’s attentiveness to a human in the absence of a reward. The experimenter commanded the dog to sit and then stood silently facing the dog. The trial started when the dog sat and ended when the dog turned away from the experimenter, capped at 2 min. The dependent measures were the amount of time that the dog remained oriented toward the experimenter on trials one and two. Based on classification in past animal literature (e.g., De Palma et al. 2005; Murphy 1998; Serpell and Hsu 2001; Vas et al. 2007), we considered this to be primarily a temperament test, measuring attentiveness and interest toward humans.

Memory problem solving

This task measured a dog’s ability to remember and efficiently recover hidden treats (similar to Barrera et al. 2015). We used the Nina Ottosson Dog Magic puzzle toy for this task. The dog watched the experimenter place treats in four equidistant wells, then cover the four baited wells along with five empty ones. The dog was given a maximum of 2 min to find the hidden treats. The dependent measures were amount of time to uncover all four treats, the number of correct wells uncovered, accuracy, and persistence. Based on classification in past animal literature (e.g., Barrera et al. 2015; MacLean et al. 2017), we considered this to be primarily a problem-solving test.

Multistep problem solving

This task measured a dog’s ability to solve a problem that required completion of two steps in a precise order (similar to Benson-Amram and Holekamp 2012; Marshall-Pescini et al. 2008). Successful performance also depended on the dog’s ability to avoid perseverating at a tempting but fruitless option (i.e., fixating on the physical location of a hidden treat despite having to manipulate the apparatus in a different location to gain access). We used the Nina Ottosson Dog Tornado puzzle game for this task. The dog watched as the experimenter baited a well, twisted the top part of the apparatus to cover the treat, and then placed a plastic bone in an empty adjacent well, thereby rendering the apparatus unable to spin until the bone was dislodged. The dog was given a maximum of 2 min to recover the treat. The dependent measures were amount of time to successfully uncover the treat, amount of time gazing to the experimenter, and amount of time perseverating (interacting with the part of the apparatus covering the baited area while the plastic bone was still in the adjacent well). We also measured persistence. Based on classification in past animal literature (e.g., Griffin and Guez 2014), we considered this to be primarily a problem-solving test.

Cylinder

This task measured a dog’s ability to inhibit perseverating at a visible but inaccessible treat directly in front of it in favor of making a temporary detour to the side to retrieve the treat (Bray et al. 2014; MacLean et al. 2014). Dogs first completed familiarization trials with an opaque cylinder,
where they learned to retrieve a treat from an open side of the apparatus without touching the front of it. In ten test trials, dogs faced an identical problem with the same solution, except that the cylinder was transparent so the food reward was readily visible throughout. The dependent measure was the number of test trials in which the dog correctly retrieved the reward on her first attempt. Based on classification in past animal literature (Hauser et al. 2002; Müller et al. 2016; Vlamings et al. 2010), we considered this to be primarily a problem-solving test.

**Detour problem solving**

This task measured a dog’s ability to detour around a barrier in order to reach a reward (similar to Fox and Stelzner 1966; Osthaus et al. 2010). The experimenter stood at the end of a Z-shaped maze and called the dog over for three trials. In the first two, she stood diagonally opposite to the dog, requiring the dog to make two turns around barriers to reach her. In the final trial, the experimenter moved directly in front of the dog but still remained behind the barrier, necessitating the same amount of detouring but with the additional challenge of inhibiting the counterproductive urge to approach directly. The dependent measures were solving times for trials one through three, as well as a test trial score (number of trials solved within the time limit). Based on classification in past animal literature (e.g., Frank and Frank 1982; Kohler 1925; Thompson and Heron 1954), we considered this to be primarily a problem-solving test.

**Greeting**

This task measured a dog’s emotional reaction to the appearance of a strange figure (similar to Goddard and Beilharz 1986). The experimenter knocked and then silently entered the room in a hunched position, draped in a felt cape. After 15 s, the experimenter then encouraged the dog to approach in a friendly tone, and patted the dog if in reach. The trial ended after 45 s. The dependent measures were latency to initially approach and amount of time spent interacting with the experimenter. Based on classification in past animal literature (e.g., Jones and Gosling 2005; Murphy 1998; Svarterberg and Forkman 2002), we considered this to be primarily a temperament test, measuring fear and confidence.

**Ball play**

This task measured a dog’s willingness to play fetch with a human (similar to Slabbert and Odendaal 1999; Wilsson and Sundgren 1998). The experimenter remained stationary and threw a ball, encouraging the dog to retrieve it. She continued throwing the ball as many times as the dog brought it back within 1 min. The dependent measure was the sum of a dog’s retrieval score over two trials. Based on classification in past animal literature (e.g., Jones and Gosling 2005; Valsecchi et al. 2011; Wilsson and Sundgren 1998), we considered this to be primarily a temperament test, measuring cooperative willingness.

**Novel object**

This task measured a dog’s emotional stability when placed alone in a room with novel objects (similar to King et al. 2003; Marshall-Pescini et al. 2017). The experimenter turned on two motion-activated toy cats, which produced erratic noises and movement, and then released the dog alone into the testing room with them for 2 min. The dependent measures were time to first approach the mechanical cats, time spent in contact with them, time spent orienting toward them, and latency to first vocalize. Based on classification in past animal literature (e.g., Barnard et al. 2016; Herrmann et al. 2007; Réale et al. 2007), we considered this to be primarily a temperament test, measuring neophobia.

**Umbrella-opening**

This task measured a dog’s initial reaction and subsequent recovery to a startling event (Sherman et al. 2015; van der Borg et al. 1991). The experimenter faced the dog and released an auto-open umbrella, which the dog was then given 45 s to explore. If the dog was not near the umbrella at predetermined time intervals, the experimenter first verbally coaxed the dog to approach, and then if needed the handler gently led the dog toward the umbrella. The dependent measures were the dog’s initial reaction score (which accounted for movement and body position), the time frame within which the dog subsequently approached the umbrella, and the amount of time spent exploring the umbrella. Based on classification in past animal literature (e.g., Foyer et al. 2014; Taylor and Mills 2006), we considered this to be primarily a temperament test, measuring reactivity and recovery.

**Program outcome**

Program outcome was coded as a binary variable: dogs either succeeded (became guide dogs or breeders) or were released (did not pass the program and were adopted by families). Breeders had to successfully complete two months of guide dog training and then were selected to enter the breeding program based on their health and behavior. Dogs could be released at any point from the program, although only 4% of our sample (n = 5) was
released prior to returning to headquarters for training. The primary reasons that dogs were released from the program were behavioral, including but not limited to lack of confidence, excitability, inability to focus, lack of initiative, body or noise sensitivity, and suspiciousness. Dogs could also be released for medical concerns or because they were transferred to an external organization. We were only interested in dogs that were released for behavioral reasons (e.g., Batt et al. 2008; Duffy and Serpell 2012). Of the 119 dogs that completed young adult testing, 20% were released for medical concerns (n = 23) or transfer (n = 1). Of the 95 remaining dogs, 63 succeeded (66%) and 32 were released due to behavioral concerns (34%) (Table S1).

Data processing and statistical analysis

All statistical analyses were carried out in R version 3.3.0 (R Development Core Team 2016).

EB or a research assistant involved in the study coded the following variables from video: perseveration during the multistep problem-solving task, initial response during the umbrella-opening task, and all variables from the isolation and novel object tasks. A separate coder then coded 20% of randomly selected trials, and interrater reliability of these variables was calculated using Spearman’s rho for continuous variables and Cohen’s Kappa for categorical variables. Interrater agreement was strong (McHugh 2012); all correlations were greater than 0.83 and the Kappa for initial reaction to umbrella-opening was 0.83. All other behavioral variables were coded live (see SI methods of electronic supplementary material).

We first examined the 34 standardized variables from the 11 tasks and used reverse coding where necessary to ensure that a positive score always indicated that the dog solved problems more effectively (e.g., more quickly) or acted in a more desirable way (e.g., confident, engaged, comfortable), whereas a negative score showed the opposite (see Table 1). We then removed those variables which did not contribute to a compact pattern of correlations, as indicated by Kaiser–Meyer–Olkin (KMO) measures of sampling adequacy that were less than 0.50 (Field et al. 2012).

Next, we used intraclass correlation coefficients estimated from variance components of a generalized linear model to determine rank-order stability of individual scores on behavioral measures between the first and second testing sessions, using the R package “rptR” (Stoffel et al. 2017). Evidence of repeatability over time is a prerequisite of temperament and cognitive traits (Brust and Guenther 2017; Carere and Locurto 2011; Guillette et al. 2017).

Then, in a hypothesis-driven approach, we applied confirmatory factor analysis (CFA), fit using the R package “lavaan” (Rosseel 2012). Tasks were placed in a factor based on how they had been previously categorized in the literature (see “General Methods” section). The “temperament” factor included tasks that measured dogs’ reactions to novel and/or startling environments, objects, and situations, their patterns of attention and distractibility, and their interactions with humans. The “problem solving” factor included tasks that measured dogs’ problem solving, memory, behavioral flexibility, and impulse control. Both factors were estimated jointly from the observed variance–covariance matrix, and parameters were estimated using maximum likelihood. Goodness of fit was evaluated using the comparative fit index (CFI: values >0.95 indicate good fit; Hooper et al. 2008), the root-mean-square error of approximation (RMSEA: values <0.08 indicate acceptable fit; Hu and Bentler 1998), and the standardized root-mean-square residual (SRMR: values <0.08 indicate acceptable fit; Hu and Bentler 1999).

Next, in a data-driven approach, we applied exploratory principal components analysis (PCA) with an orthogonal, varimax rotation to identify latent factors or attributes. We determined the number of components to retain by using parallel analysis (Horn 1965), fit using the R package “paran” (Dinno 2012), as well as the scree test (Cattell 1966) and the Comparison Data technique (Ruscio and Roche 2012).

In an effort to compare the factors obtained from the competing approaches, we looked for construct validity. First, we determined the associations between factors and questionnaire measures of temperament and impulsivity collected prior to testing. C-BARQ subscales were calculated when at least 80% or more of the values that made up the subscale were present (Duffy and Serpell 2012). Specifically, we were interested in associations with the excitability, nonsocial fear, trainability, and separation-related behavior subscales (see Appendix A of electronic supplementary material). Following Vas et al. (2007), we also calculated an activity-impulsivity score for each dog by taking the average score on items 4, 5, 6, 8, 9, 11, and 13 on the Dog-ADHD RS (see Appendix B of electronic supplementary material). All questionnaire ratings were then standardized. We used a General Estimating Equations version of a general linear regression model (GEE-GLM) to estimate all associations of interest. Because breed, sex, birth season, and age at return have been controlled for in previous studies due to their effect on behavior (e.g., Bray et al. in press), each of these was included as a covariate. We used a Gaussian error distribution with litter as the unit of analysis. Variances estimates for the statistical tests on the regression coefficients were adjusted for clustering due to litter effects using generalized estimating equations (Liang and Zeger 1993). All models were fit using the R package “geepack” (Halekoh et al. 2006). For all
models, 2-sided $p$-values <0.05 were considered statistically significant.

As a second test of construct validity, we examined the associations between factors and salivary cortisol collected upon the dogs’ return to headquarters (see above). We applied a natural log transformation to the cortisol measurements and then used GEE-GLMs as above.

Finally, to investigate the predictive validity of our factors, we used GEE-GLMs to estimate which factors were associated with success in the guide dog program. Models were the same as described above, except that they used a binomial error distribution.

**Table 1** Final set of young adult testing variables used in analyses

<table>
<thead>
<tr>
<th>Task</th>
<th>Variable</th>
<th>Description</th>
<th>Positive score equates to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolation</td>
<td>Proximity to exit</td>
<td>Time spent near the door leading out of the room</td>
<td>Less time near exit</td>
</tr>
<tr>
<td>Isolation</td>
<td>Vocal</td>
<td>Time spent vocalizing</td>
<td>Less time vocalizing</td>
</tr>
<tr>
<td>Isolation</td>
<td>Mobile</td>
<td>Time spent mobile (i.e., not sitting/laying/standing in same spot, but actively moving)</td>
<td>More mobile</td>
</tr>
<tr>
<td>Distraction</td>
<td>Toy contact</td>
<td># of toys a dog contacts while being called down the hallway, averaged over two trials</td>
<td>Less toys contacted</td>
</tr>
<tr>
<td>Sustained attention</td>
<td>Trial 1</td>
<td>Time facing the experimenter before turning away during trial 1</td>
<td>More attentive</td>
</tr>
<tr>
<td>Memory problem solving</td>
<td>Solve</td>
<td>Amount of time to successfully uncover and eat all four treats</td>
<td>Quicker to solve</td>
</tr>
<tr>
<td>Memory problem solving</td>
<td>Correct</td>
<td>Number of correct (i.e., baited) indentations uncovered in 120 s</td>
<td>More correct</td>
</tr>
<tr>
<td>Memory problem solving</td>
<td>Persist</td>
<td>Amount of time engaging with the apparatus divided by total session time</td>
<td>More persistent</td>
</tr>
<tr>
<td>Multistep problem solving</td>
<td>Solve</td>
<td>Amount of time to successfully uncover and eat treat</td>
<td>Quicker to solve</td>
</tr>
<tr>
<td>Multistep problem solving</td>
<td>Persist</td>
<td>Amount of time engaging with the apparatus divided by total session time</td>
<td>More persistent</td>
</tr>
<tr>
<td>Multistep problem solving</td>
<td>Gaze</td>
<td>Amount of time gazing at human divided by total session time</td>
<td>Less time gazing</td>
</tr>
<tr>
<td>Detour</td>
<td>Trial 2</td>
<td>Latency to solve trial 2</td>
<td>Quicker to solve</td>
</tr>
<tr>
<td>Detour</td>
<td>Trial 3</td>
<td>Latency to solve trial 3</td>
<td>Quicker to solve</td>
</tr>
<tr>
<td>Detour</td>
<td>Score</td>
<td>Number of trials solved within the time limit (1–3)</td>
<td>More trials solved</td>
</tr>
<tr>
<td>Greeting</td>
<td>Interact</td>
<td>Amount of time interacting with experimenter</td>
<td>More interaction</td>
</tr>
<tr>
<td>Retrieval</td>
<td>Score</td>
<td>Retrieval score summed over trials 1 and 2 (where 1 = no interest in ball and 5 = dog retrieves and brings back to experimenter 3+ times)</td>
<td>Better at retrieval</td>
</tr>
<tr>
<td>Novel object</td>
<td>Latency to vocalize</td>
<td>Latency to first vocalize</td>
<td>Slow to vocalize</td>
</tr>
<tr>
<td>Umbrella-opening Reaction</td>
<td>Initial reaction to umbrella-opening (where 1 = no detectable reaction and 4 = rapid avoidance response)</td>
<td>Less reactive</td>
<td></td>
</tr>
<tr>
<td>Umbrella-opening Approach</td>
<td>Latency to first approach umbrella</td>
<td>Quicker to approach</td>
<td></td>
</tr>
<tr>
<td>Umbrella-opening Contact</td>
<td>Amount of time contacting umbrella</td>
<td>More contact</td>
<td></td>
</tr>
</tbody>
</table>

**Results**

**Model fit**

We first conducted a KMO test on the 34 behavioral variables from the 11 tasks described above. The sampling adequacy was KMO = 0.52, barely above the acceptable limit of 0.5 (Field et al. 2012), so the 14 variables with individual KMO values <0.50 were removed from further consideration. Rerunning the test using 20 behavioral variables (Table 1) yielded a sampling adequacy of KMO = 0.66 and KMO values for all remaining individual variables ≥ 0.55.
We next verified that the 19 behavioral variables showed consistency across time (the novel object variable was excluded because the objects were no longer novel at the second time of testing) (Table 2). The majority of variables were strongly positively correlated across testing sessions. Thus, all subsequent analyses used data from the first testing, due to greater consistency in the time of first testing and probable ceiling effects in some tasks at the time of second testing (see SI methods of electronic supplementary material). Table S2 presents means, modes, standard deviations, and ranges of the raw scores from the first time of testing.

We first applied confirmatory factor analysis (CFA) to the 20 variables from young adult testing. Given our a priori hypothesis, we designated two latent constructs: Problem solving and Temperament (Table 3). The two factors were not correlated with one another (r = 0.08, \( p = 0.40 \)). However, results were below our threshold for acceptable fit, \( \chi^2(169, n = 119) = 509, \quad p < 0.001 \) (RMSEA = 0.130 [0.117–0.143]; CFI = 0.393; SRMR = 0.142). Given the poor fit and the fact that many of our tasks failed to load on either factor at a strength >0.32, we concluded that our first hypothesized categorization of tasks was not optimal.

We then applied exploratory PCA to the same 20 variables. Bartlett’s test, \( \chi^2(190) = 771, \quad p < 0.001 \), indicated that correlations between items were sufficiently large for PCA. Parallel analysis using 5000 iterations indicated retaining four principal components, as did inflexions in the scree plot and the comparison data technique (Ruscio and Roche 2012). The four components were orthogonally rotated. All components were uncorrelated. The total variance explained by all four components was 50%. We then examined loadings of each behavior onto the four components (Table 4), paying particular attention to loadings \( \geq 0.32 \) (Tabachnik and Fidell 2007).

Individuals who scored high on component 1 (for ease of discussion, Confident flexibility) were quick to solve a detour, even when the parameters changed slightly, were comfortable around the umbrella, mobile and exploratory during the isolation task, interactive during greeting, and proficient at playing fetch.

Individuals who scored high on component 2 (for ease of discussion, Calm + engaged) were adept at the memory problem-solving task: they solved quickly, searched in the correct locations, and appeared very engaged (i.e., persistent). They also did not have large startle responses to the umbrella.

Individuals who scored high on component 3 (for ease of discussion, Independent problem solving) exhibited superior performance on the multistep problem-solving task. They were quick to solve, highly focused on the task, and rarely looked to the experimenter.

Individuals who scored high on component 4 (for ease of discussion, Quiet investigation) rarely vocalized during the isolation and novel object tasks. They also seemed curious toward their immediate environment: they were mobile and exploratory during the isolation task, attentive to the toys during the distraction task, and focused on the experimenter during the first trial of the sustained attention task.

Thus, the exploratory PCA fit well and produced four components, all of which had behavioral loadings that combined aspects of temperament and problem solving. Taken together, the poor fit of the CFA and the superior fit of the PCA support the premise that temperament and problem solving are deeply interrelated.

### Construct validity

Table S3 presents means, modes, standard deviations, and ranges of the raw questionnaire ratings for the relevant C-BARQ and Dog-ADHD RS items.

To test our predictions regarding construct validity, we developed separate GEE-GLM models with the dog’s
questionnaire rating or salivary cortisol concentration as the predictor and individual factors (CFA- or PCA-derived) as the dependent variable. Models included breed, birth season, sex of puppy, and age at return as covariates, and assumed outcomes were correlated within litter. Results are summarized in Table 5.

The CFA factor *Problem solving* was positively associated with C-BARQ trainability (estimate = 0.08, Wald = 9.93, *p* = 0.002): dogs rated by their puppy raisers as quick learners, obedient, and attentive to their owner did well on tasks that were cognitive in nature. Furthermore, *Problem solving* was negatively associated with activity-impulsivity (estimate = −0.23, Wald = 9.73, *p* = 0.002): dogs rated as very active and impulsive by their puppy raisers did poorly on cognitive tasks. All hypothesized associations between questionnaire measures and the CFA factor temperament failed to reach statistical significance. Therefore, using puppy-raiser ratings, we found only partial support for construct validity of our CFA factors—*Problem solving* was associated with trainability and impulsivity, but *Temperament* was not associated with excitability, non-social fear, or separation, as we would have expected.

We found stronger support using puppy-raiser ratings for construct validity within our PCA factors: three of the four were associated with at least one questionnaire measure. The PCA factor *Calm + engaged* was negatively associated with excitability (estimate = −0.18, Wald = 12.28, *p* < 0.001): dogs rated as more excitable on the C-BARQ performed poorly on the memory problem-solving task and reacted strongly to the umbrella-opening. The PCA factor *Quiet investigation* was also negatively associated with excitability (estimate = −0.13, Wald = 7.70, *p* = 0.006): dogs high in excitability were vocal during the isolation and novel object tasks and curious about their immediate surroundings in the distraction and sustained attention tasks. The PCA factor *Independent problem solving* was negatively associated with both separation-related behaviors (estimate = −0.32, Wald = 8.36, *p* = 0.004) and activity-impulsivity (estimate = −0.23, Wald = 15.13, *p* < 0.001). Dogs low in separation anxiety and impulsivity were most effective at solving a multistep problem, persisting at the task and not looking to the experimenter.

As a second way of assessing construct validity, we examined salivary cortisol concentrations. Here, as predicted, *Temperament* was negatively associated with high levels of salivary cortisol (estimate = −0.26, Wald = 9.90, *p* = 0.002). In other words, dogs that scored low on *Temperament*—exhibiting anxious behaviors when left alone and wariness of the umbrella—had higher cortisol concentrations. By this measure, *Temperament* therefore had construct validity.

The PCA factor *Confident flexibility* was also negatively associated with salivary cortisol levels (estimate = −0.14,

<table>
<thead>
<tr>
<th>Observed variable</th>
<th>Latent construct</th>
<th>β</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory problem solving solve</td>
<td>Problem solving</td>
<td>0.76</td>
<td>0.08</td>
</tr>
<tr>
<td>Memory problem solving correct</td>
<td>Problem solving</td>
<td>0.73</td>
<td>0.08</td>
</tr>
<tr>
<td>Multistep problem solving solve</td>
<td>Problem solving</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Detour trial 2</td>
<td>Problem solving</td>
<td>0.18</td>
<td>0.10</td>
</tr>
<tr>
<td>Detour trial 3</td>
<td>Problem solving</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Detour score</td>
<td>Problem solving</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Memory problem solving persist</td>
<td>Problem solving</td>
<td>0.88</td>
<td>0.08</td>
</tr>
<tr>
<td>Multistep problem solving persist</td>
<td>Problem solving</td>
<td>0.26</td>
<td>0.10</td>
</tr>
<tr>
<td>Multistep problem solving gaze</td>
<td>Problem solving</td>
<td>0.21</td>
<td>0.09</td>
</tr>
<tr>
<td>Isolation proximity to exit</td>
<td>Temperament</td>
<td>0.79</td>
<td>0.10</td>
</tr>
<tr>
<td>Isolation vocal</td>
<td>Temperament</td>
<td>0.51</td>
<td>0.10</td>
</tr>
<tr>
<td>Isolation mobile</td>
<td>Temperament</td>
<td>0.68</td>
<td>0.10</td>
</tr>
<tr>
<td>Distraction toy contact</td>
<td>Temperament</td>
<td>−0.19</td>
<td>0.11</td>
</tr>
<tr>
<td>Sustained attention trial 1</td>
<td>Temperament</td>
<td>0.17</td>
<td>0.10</td>
</tr>
<tr>
<td>Greeting interact</td>
<td>Temperament</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>Retrieval score</td>
<td>Temperament</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>Novel object latency to vocalize</td>
<td>Temperament</td>
<td>0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>Umbrella-opening reaction</td>
<td>Temperament</td>
<td>−0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>Umbrella-opening approach</td>
<td>Temperament</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>Umbrella-opening contact</td>
<td>Temperament</td>
<td>0.32</td>
<td>0.10</td>
</tr>
</tbody>
</table>

*n* = 119 dogs
Wald = 4.21, \( p = 0.04 \)). Dogs with higher cortisol concentrations ended up being less proficient at solving the detour task and wary during multiple tests of temperament. Quiet investigation tended toward a negative association with salivary cortisol as well (estimate = -0.18, Wald = 3.17, \( p = 0.075 \)). Thus, when using association with cortisol to distinguish between the Temperament/Problem solving factors and the PCA factors, the data were less conclusive as there was evidence for construct validity for both.

Predictive validity

To determine the predictive ability of the different factors, we built separate GEE-GLM models for the association between each factor and program outcome. Models included breed, birth season, sex of puppy, and age at return as covariates, and assumed outcomes were correlated within litter. Results are summarized in Table 5.

Importantly, neither Problem solving nor Temperament were significantly associated with outcome in the program. Thus, neither of the CFA factors demonstrated predictive validity, once again suggesting that results do not support a clear distinction between tests of problem solving and temperament.

In contrast, the exploratory PCA produced two factors that were associated with outcome. Confident flexibility (estimate = -0.36, Wald = 4.57, \( p = 0.03 \)) was negatively associated with release from the program, indicating that dogs that scored less favorably on the detour, greeting, umbrella, isolation, and retrieval tasks were more likely to be released (Odds Ratio (OR) 0.70). We also found a significant association for Independent problem solving (estimate = -0.51, Wald = 6.07, \( p = 0.01 \)): dogs that were unengaged and performed poorly on the multistep problem-solving task were more likely to be released (OR 0.60). The other two PCA components, Calm + engaged and Quiet investigation, were not significantly associated with program outcome. In sum, when taken as a whole, the PCA factors demonstrated more evidence for both construct and predictive validity than did the CFA factors.

Discussion

Based on their assumed underlying differences and separation in much of the literature, we posited that putative temperament and problem-solving characteristics might correlate more among themselves than between each other.
When we conducted a CFA in which we forced the temperament and problem-solving variables to load onto separate factors, we found some evidence for construct validity: Problem solving was associated with questionnaire measures of trainability and impulsivity, and Temperament was associated with salivary cortisol concentrations. However, the CFA factors fell short on all other criteria. They fit the data poorly, with some tasks loading at such low levels that they were not represented by either factor. Furthermore, both CFA factors were not significantly associated with outcome in the guide dog program.

By contrast, the exploratory PCA was an open-ended analysis that searched for the combination of measures into principal components that best accounted for variation in the data. Presumably, if there had been two underlying components (temperament and problem solving) each giving rise to observed measures of its trait, the exploratory PCA would have identified them, but it did not. Rather, when we conducted exploratory PCA with no imposed structure, variables naturally sorted into four factors, each representing a different type of competence that included both problem-solving and temperament measures: Confident flexibility, Calm + engaged, Independent problem solving, and Quiet investigation. These factors fit the data well and were associated with related questionnaire and endocrine measures. Moreover, two of the factors, Confident flexibility and Independent problem solving, were significantly associated with success in the guide dog program.

Taking all the evidence together, we conclude that a specified dichotomy of problem solving versus temperament is a less informative approach to describing our data. Rather, a bottom-up method that blends features of problem solving and temperament offers a more accurate picture of our results. Of course, one reason that our bottom-up factors fit the data better than a prespecified, two-factor dichotomy is that they were derived from this particular sample. The best test of the validity of our four principal components will come when they are applied to a completely new sample. For the moment, however, our results strongly support the hypothesis that temperament and cognition are difficult to disentangle.

For example, we found that a temperament trait (separation anxiety, as evaluated through the C-BARQ) was associated with Independent problem solving, a factor consisting mainly of skills that have traditionally been conceptualized as cognitive (i.e., innovation and impulse control). Additionally, both problem-solving (detour skills) and temperament (e.g., approach and exploratory behavior)
variables loaded strongly onto Confident flexibility. Similarly, both cognitive (memory problem solving) and temperament (reactivity) variables loaded strongly onto Calm + engaged. Finally, problem-solving ability, largely regarded as a cognitive skill, loaded together with persistence, which is usually associated with temperament, on two of our factors (Calm + engaged and Independent problem solving), corroborating findings from meerkats (Thornton and Samson 2012), hyenas (Benson-Amram and Holekamp 2012), and birds (Lermite et al. 2017).

Surprisingly, none of our factors that involved bold and curious behaviors (Temperament, Confident flexibility, or Quiet investigation) were associated with ratings of nonsocial fear. We think one possible explanation for this inconsistency was that levels of nonsocial fear were quite low in our population (mean = 0.31), with very little interindividual variation (Table S3).

Previous large-scale studies that have suggested that animal cognition and temperament are separate, stable constructs have typically examined only one of these putative traits. For example, one study in birds reported correlations between three cognitive measures: innovative propensity, initial learning of a discrimination task, and reversal learning of the same task (Griffin et al. 2013). However, traits such as neophilia were not considered. Other studies in dogs (Arden and Adams 2016), mice (Matzel and Kolata 2010), and male bowerbirds (Isden et al. 2013) found evidence for a general cognitive ability that emerged across performance on multiple cognitive tests, but again did not consider the relation of these measures to measures of temperament.

Tests of temperament have also often failed to incorporate measures of problem solving into their analyses. For example, a meta-analysis of personality in dogs suggested that traits associated with temperament were moderately consistent over time (average effect size $r = 0.43$) (Fratkin et al. 2013), but the analysis did not also include measures of cognition. Similarly, a growing literature points to the existence of “behavioral syndromes,” or correlated temperament traits (e.g., boldness and aggression), across varying contexts in species as diverse as crickets (Kortet and Hedrick 2007), spiders (Johnson and Sih 2005), fish (Bell and Sih 2007), and birds (Verbeek et al. 1996). Taken together, these studies could be interpreted as implying that cognition and temperament are distinct, reliable constructs in animals. Importantly, however, none of these studies tested both purported temperament and cognitive tasks.

Another data argue against the implicit assumption that cognition and temperament are easily separable. For example, even within a single cognitive domain, like inhibitory control, context affects behavior. As a result, individual tasks meant to measure the same underlying construct are not always correlated (Bray et al. 2014; Brucks et al. 2017a; Fagnani et al. 2016). Similar issues have arisen in the temperament literature. In a study of wild baboons, Carter et al. (2012b) determined that responses to two different stimuli (a threatening stimulus and a novel object) designed to measure “boldness” were not related.

Many other studies acknowledge that cognition and temperament are difficult to disassociate, and perhaps even interdependent (Griffin et al. 2015; Sih and Del Giudice 2012). For instance, while innovation involves the cognitive ability to derive a solution, it also crucially hinges on aspects of temperament like neophilia and exploration (Kaulfuß and Mills 2008; Marshall-Pescini et al. 2017; Reader 2003). Also, numerous smaller-scale studies have successfully uncovered correlations between personality traits and problem solving across taxa (Brust and Guenther 2017; Lermite et al. 2017; Matzel et al. 2017; Nawroth et al. 2017; Schneider et al. 1991; Trompf and Brown 2014).

Two of our PCA factors, Confident flexibility and Independent problem solving, had predictive validity: scoring high on these components was associated with success in the guide dog program. These results align with prior findings. In a sample of 1067 prospective Seeing Eye® dogs, Serpell and Hsu (2001) reported that lack of confidence was the second most common reason for release from the program. Furthermore, a recent analysis of the same sample used in the current study also found superior performance on the multistep problem-solving task was associated with success (Bray et al. in press). Bray et al. (in press) additionally found a strong connection between latency to vocalize during the novel object task and outcome. However, in the current analysis, latency to vocalize loaded onto Quiet investigation, a factor that was not associated with outcome. One possibility for this discrepancy is that the explanatory power of that measure was diluted by the other behaviors that loaded onto this component.

In conclusion, psychological research has a long history of using specific tasks to study an animal’s “temperament” or “problem solving” independently, and yet empirical work rarely tests the associations between the two. We aimed to address this deficit by employing a larger-than-usual sample, a battery of both temperament and problem-solving tasks, and robust statistical methods. Our results indicate that forcing tasks into categories assumed to be associated solely with “problem solving” or “temperament” leads to poor fit and factors that do not correlate with other important variables or predict outcome. Alternatively, when a bottom-up approach allows problem solving and temperament to intermingle, the resulting factors reveal that the two domains interact in interesting ways. Furthermore, these factors are associated with questionnaire measures, cortisol, and program outcome.
Thus, by including multiple measures of both temperament and problem-solving skills, future large-scale studies of behavior could provide a more complete and realistic picture while allowing for further elucidation of the ways in which the two domains interact.

Acknowledgements Thanks to S. Bartner, L. Cohen, S. Frommer, M. Ream, and R. Schwartz for dedicating their time and expertise to behavioral testing and video coding. Thanks to Z. Cohen and I. Schamberg for their valuable feedback on drafts. We would also like to thank J. Parker and W. Lundskow from Salimetrics and K. Henning from the Institute for Interdisciplinary Salivary Bioscience Research at Arizona State University for their advice and organization of saliva collection and analysis. We are appreciative of Dr. Dolores Holle for coordinating our work at The Seeing Eye, along with Director of Canine Development Peggy Gibbon and the leadership team of The Seeing Eye. We are also grateful to the training kennel staff for giving us access to their dogs and testing space. Finally, we want to thank the puppy raising families for generously filling out our surveys.

Funding This work was supported in part by the University of Pennsylvania Department of Psychology’s Norman Anderson Graduate Student Fund, a University of Pennsylvania University Research Fund award, the Class of 1971 Robert J. Holtz Endowed Fund for Undergraduate Research, the University of Pennsylvania’s University Scholars program, and a National Science Foundation Graduate Research Fellowship to EB (DGE-1321851).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval All applicable international, national, and institutional guidelines for the care and use of animals were followed. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted.

References

Hennessey MB, Voith VL, Mazzei SJ, Buttram J, Miller DD, Linden F (2001) Behavior and cortisol levels of dogs in a public animal shelter, and an exploration of the ability of these measures to predict problem behavior after adoption. Appl Anim Behav Sci 73(3):217–237
does not correlate with mating success. Anim Behav 86(4):829–838
MacLean EL, Herrmann E, Suchindran S, Hare B (2017) Individual differences in cooperative communicative skills are more similar between dogs and humans than chimpanzees. Anim Behav 126:41–51
Pavlov I (1935) Obschchie tipy vysheh nernoj deyatel’nosti zhivotnykh i cheloveka (General types of higher nervous activity in animals and man). Poslednie soobshcheniya po fiziologii i patologii vysheh nernoj deyatel’nosti (Latest Reports on the Physiology and Pathology of Higher Nervous Activity). Akademiya Nauk SSSR, Moscow (3) 5–41