Accepted Manuscript

Test retest reliability and predictive validity of a juvenile guide dog behavior test

Naomi Harvey, Peter Craigon, Rebecca Sommerville, Caroline McMillan, Martin Green, Gary England, Lucy Asher

PII: S1558-7878(15)00155-0

DOI: 10.1016/j.jveb.2015.09.005

Reference: JVEB 919

To appear in: *Journal of Veterinary Behavior*

Received Date: 12 September 2014

Revised Date: 18 September 2015

Accepted Date: 21 September 2015

Please cite this article as: Harvey, N., Craigon, P., Sommerville, R., McMillan, C., Green, M., England, G., Asher, L., Test retest reliability and predictive validity of a juvenile guide dog behavior test, *Journal of Veterinary Behavior* (2015), doi: 10.1016/j.jveb.2015.09.005.

Test retest reliability and predictive validity of a juvenile guide dog behavior test

Harvey, Naomi*[1]; Craigon, Peter[1]; Sommerville, Rebecca[2]; McMillan, Caroline[1]; Green, Martin[1]; England Gary[1]; and Asher, Lucy[3]

[1]The University of Nottingham, School of Veterinary Science & Medicine, Sutton Bonington, LE12 5RD, UK.

[2]Royal School of Veterinary Studies, Easter Bush, Midlothian, Edinburgh, EH25 9RG.

[3]Centre for Behavior and Evolution, Henry Wellcome Building, Newcastle University, Newcastle, NE2 4HH, UK.

*Corresponding author: Naomi.Harvey@nottingham.ac.uk

Keywords: Juvenile; Guide dog; Behavior test; Consistency; Validity; Reliability

ABSTRACT

The ability to measure stable and consistent behavioral traits in dogs would facilitate selection and assessment of working dogs, such as guide dogs. Ideally, these measures should predict suitability for the working role from a young age. This study assessed test-retest reliability of a juvenile guide dog behavior test and predictive validity using qualification or withdrawal from guide dog training. Ninety-three guide dog puppies (52F; 41M) were tested at 5 (mean 4.78; ± 0.73 SD) and 8 (mean 7.98; ± 0.78 SD) months of age. The dogs were exposed to a sequence of 11 stimuli designed to assess the dogs' reactions to: meeting a stranger, obedience commands, body sensitivity, scavenging, and 'animal' and human distractions. The behavior of dogs was digitally recorded and analysed using an ethogram incorporating both frequency of behavior and specific reactions to stimuli. Test-retest reliability indicated inter-individual consistency in many of the behavioral measures such as jumping, barking and 'low' greeting posture. Behavior measures that did not show inter-individual consistency between tests included obedience responses, lip-licking, body shaking and scratching. Binary logistic regression models revealed seven behavioral measures at five months and five measures at eight months that were significantly associated with qualification or withdrawal. Uncorrelated measures and principal component scores of correlated measures were combined in a logistic regression model

that showed great potential for predicting the probability of a dog qualifying or being withdrawn from guide dog training.

INTRODUCTION

Puppy testing (the assessment of behavioral responses in puppies) has been referred to as "the holy grail of temperament testing" (Miklosi, 2011). This description reflects the potential value of predicting future behavior from a young age to future owners and rehoming and working dog organisations. Valid and reliable behavior tests could be invaluable, enabling the selection of dogs suitable to owner need, specific working dog roles such as support, police or guide dogs, and aid in suitable placement of puppies to homes from rescue shelters (King et al., 2012). The four periods of development during 'puppyhood' are the neonatal, transitional, socialisation, and juvenile periods (Scott and Fuller, 1965). The juvenile period is the longest, beginning at approximately three months of age and continuing until sexual maturity (Scott and Fuller, 1965). Domestic dogs typically undergo sexual maturity between 6-9 months of age, but behavioral, or social, maturity is considered to be achieved anywhere from 12 to 24 months of age depending on breed (Overall, 2013). Despite the juvenile period being defined as ending at sexual maturity, the majority of published studies consider dogs less than 1 year of age to be puppies and dogs greater than 1 year of age to be adults, or young adults (Fratkin et al., 2013).

The juvenile period is currently the least studied or documented stage of puppy development. The majority of what is known about neural and behavioral development in the dog focuses on the first 8 to 12 weeks of life (Scott and Fuller, 1965) and little is known about what further changes may occur in regards to neural development after 12 weeks (Overall, 2013). However, evidence from human and rat studies show that the mammalian neural network continues to grow and develop throughout adolescence and that this can have long term effects on adult personality (McCrae et al., 2000; Sisk and Zehr, 2005; Crone, 2009; McCormick and Mathews, 2010).

While some studies have shown associations between puppy test results, and training outcomes of adult working dogs (Slabbert and Odendaal, 1999; Svobodova et al., 2008; Asher et al., 2013), the majority of previously developed puppy tests have had limited to no success in predicting adult behavior (Goddard and Beilharz, 1986; Wilsson and Sundgren,

1997b; Wilsson and Sundgren, 1998a; Wilsson and Sundgren, 1998b; Riemer, et al., 2014). However, these tests were mainly conducted on dogs in the early stages of development, below 12 weeks of age. The lack of success in predicting adult behavior shown by many puppy tests could be explained by continuing neural and behavioral changes within juvenile dogs, which are likely to continue past sexual maturity, stabilizing at only social maturity (Overall, 2013). This is supported by evidence which shows that the predictive ability of behavior tests improve as an animal ages (Goddard and Beilharz, 1986; Wilsson and Sundgren, 1998a; McCrae et al., 2000; Hoffmann, 2002; Bell et al., 2009; Fratkin et al., 2013). Therefore, conducting assessments on juvenile or young adult dogs, rather than dogs less than 12 weeks of age, could improve a tests predictive value.

Previous research indicates that behavior in juvenile and young adult dogs, aged as young as 5 months, can be partly predictive adult behavior (Hoffmann, 2002; Duffy and Serpell, 2012). Significant associations were found between suitability to the guiding role and scores on a questionnaire known as the C-BARQ, when completed by volunteer puppy carers (known as puppy walkers or puppy raisers) about behavior of dogs' aged 6 and 12 months (Duffy and Serpell, 2012). While the results of Duffy and Serpell indicate that prediction of working suitability could be possible from 6 months of age, the questionnaire scores were unable to actually separate individual dogs that went on to qualify or be withdrawn, and so could not be used to categorically predict the training outcome of a given individual.

There may be further applications for predicting adult behavior on the basis of canine personality. Stable and consistent differences in behavior have been demonstrated for dogs less than 1 year of age (Fratkin et al., 2013). The mean correlation for personality tests assessed at different ages and across a large number of studies was 0.34, similar to human behavioral consistency measures (Mischel, 2006). Yet the majority of published studies that assess juvenile dog behavior (Goddard and Beilharz, 1986; Wilsson and Sundgren, 1998a; Wilsson and Sundgren, 1998b; Slabbert and Odendaal, 1999; Batt et al., 2008; Sforzini et al., 2009; Kim et al., 2010; Duffy and Serpell, 2012; Asher et al., 2013) fail to provide evidence of the test's reliability, have been conducted on too few dogs for the results to be meaningfully interpreted (Batt et al., 2008; Sforzini et al., 2009; Kim et al., 2010). For any behavior to be 'predictive' of a future event or outcome, it must be reliable, consistently recorded, and consistent in performance over time (Diederich and Giffroy, 2006; Taylor and Mills, 2006).

Personality traits must be consistent and stable over time (McCrae et al., 2000; Uher, 2011), so tests that have been shown to predict the same behavior at a later date may be measuring aspects of personality. Questionnaire based assessments can be used to assess dog personality (e.g. Duffy and Serpell, 2012), but the most commonly employed method is the test battery (Jones and Gosling, 2005). Test battery approaches using ethograms can be used to assess 'personality' (Sinn et al., 2010; Wilsson and Sinn, 2012; Fratkin et al., 2013), and are considered less subjective than questionnaire assessments. Tests are conducted under controlled or semi-controlled conditions and involve exposing dogs to a series of stimuli while recording behavior either at the time, or subsequently from video footage (Highfill et al., 2010; Carter et al., 2012; Wilsson and Sinn, 2012). Scoring protocols associated with practical behavior tests are reductionist in nature, breaking down complex series' of behavior into small constituent parts that can fail to capture subtle or rare behavior (Asher et al., 2009; Uher, 2011). Test batteries are often employed by staff in rescue shelters who wish to evaluate a dog's behavior to aid in successful rehoming, and in decisions regarding euthanasia or rehabilitation (Dowling-Guyer et al., 2011; Mornement et al., 2014,) as well as in working dog organisations that use military dogs (e.g. Haverbeke et al., 2009), police dogs (Slabbert and Odendaal, 1999), or run breeding programs (Arvelius et al., 2014).

Predictive validity of behaviour tests could also be improved by ensuring that the situations under which the tests are conducted, and the stimuli encountered, closely reflect the situations to which the results are meant to be applied (Taylor and Mills, 2006; King et al., 2012; Mornement et al., 2014). Two tests of shelter dog behaviour, which provided sufficient evidence of test reliability, and have successfully predicted future behaviour of dogs following rehoming were both designed to reflect everyday situations, often conducted in the dogs home kennel (Dowling-Guyer et al., 2011; Valsecchi et al., 2011; Marder et al., 2013). It is possible that the novel stimuli encountered under artificial testing situations may make the tests inherently stressful for the subjects, reducing the range of traits that can be studied to those related to stress or anxiety, and weakening the validity of the results (Rayment et al., 2015).

The main aim of this study was to design and evaluate a test battery for juvenile dog behavior using a behavioral coding ethogram, for predicting outcomes in a guide dog

training programme. A subsidiary and related aim was to investigate which aspects of behavior measured in the test were consistent and stable over time and so could be related to personality. To achieve these aims we assessed: 1) test-retest reliability (temporal consistency) between tests at two different ages; and 2) predictive criterion validity by comparing dogs test scores to their outcome within the Guide Dogs' training program (qualification as a guide dog or withdrawal from the program for behavioral reasons).

METHODS

SUBJECTS

The target population was defined as all Guide Dogs' puppies born in December & January 2011 who were tested once at 5 months and again at 8 months of age. Potential guide dogs are cared for by volunteer 'puppy walkers' (PWs) during the formative months of their life. Contact details of all volunteer puppy walkers, nationwide, due to receive these puppies were obtained (n=148). A postcode map of participant locations was created using online mapping software Batchgeo (http://batchgeo.com/). Puppy walkers whose locations were more than a two-hour drive from another puppy walker were removed from the study sample. The remaining 119 PWs were invited by letter to participate with their dog. The 93 PWs who consented to participate met with the researchers at the venue closest to them (see below). PWs were briefed over the phone, and by letter, on the content of the test battery.

Ninety-three dog-PW dyads participated in the study (69 tested twice, 13 tested only at 5 months, and 11 tested only at 8 months). The mean age of dogs tested in the first test was 4.78 months (±0.73 SD); and in the second test was 7.98 months (±0.78 SD). Of the 93 dogs tested, 52 were female and 41 male (first test 48F/34M; second test 44F/36M). The dogs came from 29 litters, with 23 different sires. The dogs tested (sire x dam) were 39 golden x Labrador retrievers; 38 Labrador retrievers; 8 Labrador x golden retrievers; 6 Labrador x golden retriever crossbreeds; and 1 German shepherd x golden retriever.

TEST ARENA

Tests were conducted in 21 different venues, typically village halls, church halls or community centres with at least two rooms, one of a minimum size of 7m by 5m for testing and another for use as a waiting area. All venues also had an outside space not adjacent to a road. The test battery was named the 'juvenile guide dog behavior test'.

A test arena, measuring 6.5 x 4.5 metres, was marked out at each venue using rows of chairs, and always included an entry/exit route in view of at least one camera (Figure 1). Video recordings of the indoor test arena were made using three camera's (Camera 1 was a Panasonic HDC-HS60; Camera's 2 and 3 were wide angle GoPro HD-Hero2) mounted on chairs. A pathway in an outside area, which measured a minimum of 14 metres in length, was established with stimuli consistently placed at a measured distance from the path (Figure 2). Filming of the outside area was permitted by the use of a head mounted camera on Experimenter 1 (wide angle GoPro HD-Hero2) positioned at approximately a 45^{o} downward angle.

PROCEDURE

The test procedure was developed following an extensive review of literature, consultation with Guide Dogs' training staff, three months of observations of puppy behavior in Guide Dogs' puppy classes, and pilot work with juvenile pet and potential guide dogs. Subtests were designed to address behavior that could be representative of distractions (from food, animals or people), training and obedience, and body sensitivity. The food distraction subtest was designed to replicate situations where food rubbish is encountered on walks, which is problematic in guide dogs (Murphy, 1998). No stimuli or procedures were considered which had the potential to induce a strong fear response. To maximise the test's validity, efforts were made to make the protocol as 'normal' and stress-free as possible for the dogs by mimicking situations they could encounter on a day-to-day basis. Testing took place during the months of May-June and August-September 2012.

A total of 11 subsets were used: 1) Meet a stranger; 2) Obedience with PW; 3) Obedience with stranger; 4) Raised path; 5) Body check; 6) Head ring; 7) Tea-towel; 8) Food; 9) Robin; 10) Pigeons; and 11) Human distraction (see Table 1). Two subtests, 1 & 5, were adapted from subtests 1 and 3 from the 'Social Contact' task in Svartberg (2005). Three

experimenters were involved and the main handler for the tests, Experimenter 1 was kept out of sight from the dogs until the test began.

Equipment for subtests 1-7 included two polyethylene foam blocks (L600mm, W400mm, D80mm), placed end-to-end to form a raised path for subtest 4, sourced from Foam Solutions UK (http://www.foamsolutionsuk.co.uk), a rubber 13" Aerobie® Pro Ring for subtest 6, and a quarter folded cotton tea-towel for subtest 7. Drawstring treat bags were worn by experimenters 1 and 2 clipped onto their belts that contained a mixture of two types of dog treats (Misfits®: Ruff Rips™ and Scruffy Bites™). Equipment for subtests 8-11 consisted of two small cones used to mark the beginning of subtest 8, two paper plates holding three torn up hot dog sausages (Herta® Frankfurters Classics), two plastic, whole pigeon decoys with legs (head down) (www.countrykeeper.net), and an RSPB 'singing' robin tied to a pulling device). The pulling device consisted of an adjusted remote control car with a retractable dog lead joined to its wheel, hidden in a cardboard box by a woollen blanket. The car was activated by remote control and the lead then pulled the robin into a second cardboard box 'hide'.

VIDEO ANALYSIS

An ethogram of behavioral responses was created prior to behavioral testing (Table 2). A single rater scored all videos over a five-month period.

STATISTICAL ANALYSIS

Tests for correlations, associations between variables and principal components analysis were undertaken using SPSS v. 21 (SPSS Inc., Chicago, IL, USA). Logistic regression analysis was undertaken in R version 3.0.2 (R Core Team, 2013); R scripts available on request. Unless otherwise stated, significance was set at $P<0.05$.

INTRA-RATER RELIABILITY

Intra-rater reliability was assessed for all ethogram measures. For measures that were repeated, due to subtest replicates, they were combined so that just the measure was assessed. For example, tail height was recorded for the two replicates of subtest 6 as '1st Tail

height' and '2nd Tail height' but for this analyses the replicates were combined to give just 'Tail height'. Cohen's Kappa (*K*) was utilized to assess binary data (Gwet, 2014), and intraclass correlation coefficients (ICC's) were calculated for continuous data using a two-way mixed model with consistency (Nichols, 1998). Mean weighted kappa coefficients are most commonly used to assess agreement for ordinal data where there is an underlying continuum (Roberts and McNamee, 2005). Average measures ICC's with absolute agreement are directly equivalent to the mean weighted kappa (Fleiss and Cohen, 1973), so average ICC's were applied to ordinal data. Cohen's Kappa (*K*) is most often interpreted as follows: less than 0.20 is poor, unacceptable correlation, 0.21-0.4 is a fair and acceptable correlation, 0.41-0.60 is moderate correlation, 0.61-0.80 is a good correlation, and 0.81-1.00 a very good correlation (Altman, 1991). Guidelines for interpretation of both mean weighted kappa and ICC coefficients suggest that below 0.40 is poor or unacceptable, between 0.40-0.59 is fair, between 0.60-0.74 is good, and above 0.75 is excellent (Cicchetti, 1994; Bryington at al., 2004). ICC coefficients of above 0.60 were considered acceptable for this analysis.

Using methods outlined by Walter et al. (1998), a sample size estimation based upon α=0.05, β=0.20, with a minimum acceptable coefficient of 0.60 and a maximum expected coefficient of 0.80, provided an acceptable sample size of 39.1. Further sample size guidelines for intra-rater reliability of tests using ICC statistics suggest that 40 samples with 2 replicates are sufficient to obtain precise coefficients (where precision is shown by 95% confidence interval widths of less than 0.40) when the coefficient is above 0.50 (Gwet, 2014). Based upon these two guidelines, videos of 40 tests were analysed twice by the same rater, approximately two years apart.

TEST-RETEST RELIABILITY

To investigate test-retest reliability all individual measures were tested for correlations between the 5M and 8M tests. Of the 93 dogs in this study, 69 participated in both tests and form the basis of this analysis. To assess test-retest reliability (for which rank-order consistency is assessed) Kendall's tau-b was used for binary variables, and Spearman's rank for ordinal and continuous data. P values are presented with and without (for comparison

with existing literature) correction for multiple testing using the Improved Bonferroni Procedure (Simes, 1986).

PREDICTIVE VALIDITY

Of the 93 dogs tested, 61 qualified as guide dogs (Q), 22 were withdrawn for behavior reasons (W-B), 4 were withdrawn for health reasons and 6 were selected for breeding. For the purposes of this analysis only test scores of those dogs that were qualified or withdrawn for behavior reasons were be used. This gave a sample size of 73 dogs (52Q and 21W-B) with 5M test scores and 72 dogs (56Q and 16W-B) with 8M test scores.

Separate binary logistic regression analyses were conducted for the 5M and 8M tests. The basic model equation using a logit link function can be written as:

$$y_i \sim \text{Binomial}(n_i, \pi_i)$$

$$\text{logit}(\pi_i) = \log\left(\frac{\pi i}{1-\pi i}\right) = \beta_0 + \beta_n X_i$$

Where y_i represents the response variable (withdrawal for behavior vs. qualification as a guide dog) for the ith dog; π_i represents the probability that $y_i = 1$; β_0 is the model intercept (the estimated response value when the predictor equals zero), and the regression coefficient for the explanatory variables are represented by $\beta_n X_i$ (where '$_n$' indicates the variable ID).

A four-step process was utilized for multivariate analyses: (1) univariate logistic regressions were run for each variable from the test against training outcome, criteria for retention of variables was set to $p<0.1$; (2) to avoid multi-collinearity, correlations between retained variables (Spearman's for continuous measures, Kendall's tau-b for ordinal measures, McNemar's tests for binary measures, and Mann Whitney U tests to compare binary against ordinal or continuous measures) were conducted and where correlations were significant ($p<0.1$) principal component analysis (PCA) was utilized to reduce the variables by creating component scores (PCAs based on Eigen values >1, with varimax rotation), following guidelines set out by Budaev (2010) for studies with fewer than 100 subjects loading values of >0.50 were considered significant; (3) PCA scores and remaining uncorrelated variables were entered into a composite logistic regression model using a backwards elimination procedure; (4) the 'anova' command of the statistical package was then used to assist with

selection of the best fitting model. Figures were made by plotting the probability of being withdrawn, and outcome (withdrawn or qualified), against the probability of being withdrawn and a composite score (calculated from the model), which will henceforth be referred to as the "composite model score".

RESULTS

INTRA-RATER RELIABILITY

The variables 'approaches' and 'avoids' from subtests 8 to 10 showed too little variation in this random subsample of tests. Only 'approaches' for subtest 10 (Pigeons) could be assessed for intra-rater reliability. Following the combination of measures that were repeated within replicates of subtests 5 to 7, this created 33 variables for which intra-rater reliability was testable.

For the 16 testable binary variables, one could be classified as showing 'fair' agreement with a *K* of 0.35, according to Altman (1991) and one could be considered to have 'moderate' agreement, with eight showing 'good' agreement and 6 showing 'very good' agreement (Table 3).

For the 17 continuous or ordinal variables evaluated here, all showed ICC values above 0.60, with three being classified as 'good' and 14 classified as 'excellent' according to the guidelines set out by Cicchetti (1994) (Table 4).

TEST-RETEST RELIABILITY

Most behavioral measures considered showed some temporal, test-retest consistency between 5 and 8 months (Table 5). Twenty-five measures were significant before correction for multiple testing, with 18 remaining significant after correction. Measures that did not show temporal consistency include: shakes, scratches (across subtests), pull strength (in subtest 1), the response to sit, wait and down commands, proportion of time spent gazing at the puppy walker, and lip-licks (in subtest 2 & 3), crossing a raised pathway (subtest 4), compliance score in a body check test (subtest 5), body posture in the 2nd trial of the head-ring (subtest 6), turns head towards a tea-towel placed on the back (subtest 7), time

orientated towards, approaching or avoiding a pair of fake pigeons (subtest 10), and pull strength towards an unknown person (subtest 11).

PREDICTIVE VALIDITY

FIVE MONTHS

Ten variables at 5M showed associations with qualification or withdrawal to the $p<0.1$ level (Table 6). Six were found to be significantly associated with each other and were included in a PCA, which yielded two components (Table 7). The 5M PCA achieved a KMO statistic of 0.59 and Bartlett's test of sphericity of $p<0.05$. Three variables loaded strongly in the first 5M component (5M-PC1), all of which came from subtest 6 (tea-towel). High 5M-PC1 scores were achieved by dogs that attempted to remove the tea towel on both repetitions and played with it on the second repetition. The remaining three variables loaded on 5M-PC2, and high scores for this component were indicated by barking at any point during the test, lip-licking in subtest 2 (PW obedience) and not-shaking for subtests 5-7 (body sensitivity tests). A significant composite logistic regression model could be formed for the 5M test combining each component score (5M-PC1 and 5M-PC2) with three independent variables: time oriented towards the food in subtest 8, and 'Down' performance in subtest 2 and subtest 3 ($Z=3.81$, $p<0.001$, $R^2=48.4\%$, Figure 3). For each 1 unit increase in the composite score the odds of being withdrawn for behavioral reasons increased by x1.7 (95% CI 1.63 to 4.55).

EIGHT MONTHS

Ten variables at 8M were associated with qualification or withdrawal to the $p<0.1$ level (Table 6). Nine of these variables were significantly associated with each other and so were included in a PCA, which yielded three components (Table 8). The 8M PCA achieved a KMO statistic of 0.65 and Bartlett's test of sphericity of $p<0.05$. The first component, 8M-PC1, contained mostly variables from subtests 8-10 (distraction circuit), dogs with high scores on 8M-PC1 pulled more strongly towards the food, Robin and Pigeons, and played with the tea-towel the first time in subtest 7. Component two, 8M-PC2, contained two variables; dogs with high scores on this component avoided the Pigeons and showed a change from neutral posture in the first repetition of subtest 7 (tea-towel). Component three, 8M-PC3, contained

two variables loading above 0.50 (turning to look at the tea-towel in subtest 7 and approaching the robin in subtest 9), and one variable with a loading on 0.47 (time oriented toward the food).

A significant composite logistic regression model could be formed for the 8M test combining each component score (8M-PC1, 8M-PC2, 8M-PC3) with the one independent variable: 'low' greeting posture in subtest 1 ($Z=3.64$, $p<0.001$, $R^2= 52.3\%$, Figure 3). For each 1-unit increase in the composite score the odds of a dog being withdrawn increased by x1.7 (95%CI 1.59 to 4.66).

DISCUSSION

The aim of this study was to design a battery of practical tests for assessing juvenile dog behavior using a behavioral coding ethogram, which would record and identify behavior associated with the dog's personality and had potential to predict suitability for the guiding role. The study found evidence for both reliability and validity of this test (Taylor and Mills, 2006, Martin and Bateson, 2007). To identify behavior that may be indicative of personality, reliable measurement items were assessed for test-retest reliability (temporal consistency). Temporal consistency was good with a mean correlation of 0.41 for 25 measures, and 0.45 for the 18 measures that remained significant after correction for multiple testing. These results compare favourably to published literature, which together showed a mean correlation of 0.34 (Fratkin et al., 2013). To assess validity, we considered the association between measurements and outcome in Guide Dogs' training programme, finding seven measures at five months and five measures at eight months that were significantly associated with qualification or withdrawal individually. Additionally, a logistic regression model could be produced for each age tested that demonstrated potential for identifying dogs likely to qualify or be withdrawn from the training program.

INTRA-RATER RELIABLITY

Intra-rater reliability of the ethogram revealed the majority of measures achieved good to excellent consistency, with all falling within acceptable limits of agreement (Altman, 1991,

Cicchetti, 1994). As predicted in sample size estimation the 95% confidence interval width for all ICC statistics was less than 0.40, which lends credibility and confidence to these results. It is essential to establish reliability of scoring methods, especially where decisions are made based upon their results. While intra-rater reliability has been demonstrated within this study, it will need to be re-assessed in any future application of the behavior test when new raters are trained (Martin and Bateson, 2007). Inter-rater reliability was not assessed in this study because all tests were scored by a single rater. If multiple rates are used, as is commonly the case, inter-rater reliability will also need to be demonstrated.

TEST-RETEST RELIABILITY (TEMPORAL CONSISTENCY)

The results from the test-retest analysis show that many behavioral measures from the ethogram achieved good (>0.3) to high (>0.6) correlations between the time points, suggesting the presence of inter-individual consistency. These results compare favourably to those found in a meta-analysis of behavioral consistency in dogs where similar studies on puppies (dogs <1 year old) had a mean correlation between tests of 0.34 (Fratkin et al., 2013).

Measurement items that showed poor temporal consistency in rank order of individuals were obedience task response, gaze behavior and summed counts of lip-licking, shaking and scratching. Intra-observer reliability for these measures was acceptable to good, which suggests the lack of correlations between tests is due to instability of the behavior, not recording error. These results suggest that these behavior are most subject to change, and cannot be considered behavior that directly reflect personality traits due to their lack of inter-individual consistency (Freeman et al., 2011).

Behavior that showed medium to high consistency correlations (*rho* of >0.4) across the three month time period included jumping, barking, whining, 'Low' posture upon greeting, mouthing, human licking, and ear and tail position. The measures from the distraction subtests (8-11) also showed good to high consistency, confirming that they detect consistent individual differences in behavior. These measurement items could be used as measures of dog personality. The measures from the distraction subtests were designed to

assess distraction related tendencies, but to be sure that they measure a distraction trait (e.g., Arata et al., 2010) would require comparison with independent measures.

The high level of inter-individual consistency for the distraction measures, compared to the other measures, contradicts one study that showed low repeatability for distraction measures in dogs tested at 6 and 12 months of age in an Australian guide dog population (Goddard and Beilharz, 1984). Differences between these studies may stem from differences in the test and recording methods. Our study used semi-controlled situations and objective behavioral coding methods to score the dogs, and the re-test interval was half that of Goddard and Beilharz and behavior is more consistent across shorter time intervals (Bell et al., 2009, Fratkin et al., 2013). Goddard and Beilharz (1984) observed the dogs in uncontrolled conditions and used a more subjective scoring system. Assessments of distraction behavior should be conducted under standardised, controlled or semi-controlled conditions.

Our results compare favourably with those from other test-retest studies of behavior in dogs. Sinn et al. (2010) found medium to high, significant correlations (0.4-0.6) for behavior scores between tests with short intervals (1-30 days) for US Military Working Dogs. Correlations decreased to <0.3 with longer intervals (30 - 157 days). In our study the interval between tests was approximately 91 days (13 weeks), and medium to high correlations were achieved with a mean significant correlation of 0.41.

There was a lack of correlation between the 5M and 8M tests of obedience, which suggests that obedience, itself, may not be an aspect of personality in dogs. In a meta-analysis of consistency of personality 'traits' in dogs, 'Responsiveness to Training' was found to have the lowest overall consistency of the 'traits' assessed (Fratkin et al., 2013). Such assessments of trainability are often based on questions about obedience, so it is probable terms are being used synonymously in the scientific literature. Obedience has a strong reliance upon factors external to the dog including amount, type, and quality of training, which are not often assessed in such tests and may mask dog effects.

PREDICTIVE VALIDITY

Some test measures discriminated between dogs that eventually qualified or were withdrawn for behavior, at both 5 and 8 months of age. Only one measurement item was significantly associated with the dogs' training outcome from both tests: time oriented towards the food in subtest 8. Dogs who spent longer oriented towards the food had increased chances of withdrawal from the training program.

Expression of a 'low' posture, as defined in Table 2, during greeting in subtest 1 was found to be positively associated with success in guide dog training. Low postures have been associated with the experience of both chronic and acute stress (Beerda et al., 1998; Haverbeke et al., 2009). Our definition of 'low' posture included that the dog wagged its tail. While in this position the dogs often licked the hands of the experimenter, a behavior associated with human-greeting in dogs (Westgarth et al., 2008). This version of a 'low' posture occurs only during greeting and is accompanied by tail wagging (and potentially hand licking), and could be considered to be an appeasement posture that may reflect a particularly 'sociable' dog. It is possible that dogs viewed as more 'sociable' could be more likely to qualify as a working guide dog.

Body shaking behavior was also associated with qualification from guide dog training and is also thought to be associated with the experience of anxiety or internal conflict in a dog (Beerda et al., 1997). However, in our study, shaking following the 'body sensitivity' subtests substantially decreased the odds of a dog being withdrawn. One possible explanation for this unexpected association could be that shaking is a coping behavior, expressed to help alleviate anxiety. Shaking behavior was not temporally consistent, and only shaking at 5 months was associated with a dog's training outcome. The presence of lip-licking during the puppy walker obedience subtest at 5 months was also associated with increased chances of withdrawal, and also did not show temporal consistency. In our study shaking and lip-licking were shown not to predict future shaking or lip-licking, but they did appear to represent an aspect of the dog's state at the time of testing, which was predictive of the independent event of qualification as a guide dog more than a year later.

Using composite regression models, the factors of most importance in predicting outcome at five and eight months were identified. At five months, the dogs that qualified responded

the first time to the 'down' command from their puppy walker, responded the second or third time to the novel person for the same command, and scored low on the two five month component scores. The first five-month component score included attempted removal of the tea-towel from their back in subtest 7 (on each replicate) and playing with the tea-towel in the second replicate. This component represents a subtest specific score regarding the dogs' reaction to a garment-like fabric being placed on their back. The second five month component score included barking in any subtest, lip-licking during obedience with their puppy walker, and an absence of body shaking after subtests 5-7 (body sensitivity tests). Barking, lip-licking and shaking may be associated with internal conflict or anxiety (Beerda et al., 1997). These components could contain some aspect of responses to anxiety provoking situations. If so, they may reflect behaviors defined under the 'Fearfulness/nervousness' dimension (McGarrity et al., 2015).

At eight months, the dogs that were statistically predicted as most likely to qualify as guide dogs were those which did not display a 'low' greeting posture, had low scores on the first component (distraction) and/or second component (fear/anxiety) identified from a PCA, and/or high scores on the third component (low reactivity). The first eight-month component included pulling more strongly towards the food, robin, and Pigeons from subtests 8-10 and playing with the tea-towel from the first replicate of subtest 7. This component appears to represent distraction-related behavior, one of the most common reasons for withdrawal within Guide Dogs in the UK, and other guiding schools (Arata et al., 2010). The second 8-month component included avoidance of the Pigeons from subtest 10, and change from neutral posture in response to the first tea-towel replicate in subtest 7. These behaviors may be indicative of a fearful or anxious response. Interpretation of these behaviors would be aided by concomitant assessment of physiological variables, such as heart rate or circulating glucocorticoid levels (Rayment et al., 2015). It may appear contradictory that dogs least likely to qualify as a guide dog are those that pulled harder towards the Robin and those that also avoided the robin. While dogs would be unlikely to show both behavioral responses simultaneously, strong avoidance or approach behavior with respect to novel items is undesirable for a working guide dog. The third 8-month component was based upon turning to look at the tea-towel on their back in subtest 7, and approaching the robin in subtest 9, where a lack of such behavior was associated with

qualification. Although the principal components discussed here may place dogs within proposed personality dimensions (such as 5M-PC2 within 'fearfulness/nervousness'), it is important to note the methodological limitations of this study. The main aim was to identify behavior that may be predictive of guide dog suitability, as such the principal components were formed only from behavior that showed predictive associations and cannot be considered to be exclusive measures of dog personality traits. Additionally, behavior included in the predictive models was not required to be temporally consistent in order to predict guide dog suitability. For a behavioral measure to be considered a measure of personality it must be temporally consistent. Therefore any placement of these principal components within a personality framework must be done with caution.

The composite regression models highlighted the test's ability to identify dogs with high and low probabilities of withdrawal for behavior. Models based on probability of withdrawal could be utilized as a tool to aid decision-making regarding a dog's training, or subsequent inclusion in the training program. The model was able to classify a dog's outcome (qualification or withdrawal) correctly for 79.7% of dogs for the 5M model and 87.3% for the 8M model. These values compare favourably with previous literature where 78% of adult dogs (15-18 months-old) in the Swedish Armed Forces programme were correctly classified by a behavioral coding method (Wilsson and Sinn, 2012). Our results were based on a default threshold of 50% probability of success as a guide dog to classify dogs as either likely to qualify or likely to be withdrawn. Based upon the requirements of organisations such as Guide Dogs, a highly conservative threshold for automatic withdrawal of a dog could be set at 90% probability. Dogs with a probability of withdrawal of between 60-90% could be given a 'flag' that would allow their progress to be monitored more closely and for the application of potential rescue strategies. Dogs with a probability of withdrawal of less than 10% could be fast-tracked through the system, or individuals with desired physiological phenotypes within this group could be selected for breeding. Using a 60% probability as a threshold for alerting dogs likely to be withdrawn would yield positive predictive values (correctly identified withdrawn dogs) of 55% and 50%, for the 5 and 8-month tests, with 92% and 80% of dogs scoring above the cut-off being withdrawn. Positive predictive values (PPVs) are rarely reported from behavioral assessments of working dogs, but Asher et al. (2013) noted that a puppy test for 8 week old guide dog puppies yielded an 8% PPV (Asher et al., 2013).

One test of 6 month old trainee police dogs had a 33% PPV (Slabbert and Odendaal, 1999). Positive predictive values of 50% and 55% from the test described here are high, and could be of significant value to Guide Dogs.

CONCLUSIONS

The test presented here represents a new behavior test for juvenile dogs from which reliable and consistent measurement items have been identified. Some of these measurements have shown considerable predictive criterion validity for guide dog suitability. This juvenile guide dog behavior test has the potential to be used as a decision making tool for Guide Dogs, by identifying dogs who will not be successful while they are still puppies. Identification of dogs most likely to qualify could assist with selection of dogs for inclusion in the breeding program. As with many test batteries, the application and subsequent scoring associated with this test in its current form is labour intensive compared to that of a rating style assessment. Not all elements of the test included measures shown to be of predictive value, such as subtest 4 (Path). The test order was not randomised, so as with most tests there is the potential for order effects on the dogs' behavior. The overall test length was below 20 minutes, within the minimal length suggested by Taylor and Mills (2006). If the only purpose of the test were to predict guide dog training outcome, only behavior that showed significant associations with guide dog qualification or withdrawal would need to be recorded and measured from video footage. Combined with additional assessment methods, this test could be applied to those dogs whose behavior is already under question, to gain further estimates of their chances of success in training. The juvenile guide dog behavior test and its associated ethogram could also be utilized for future scientific studies of juvenile dog personality and behavior, which has broad applicability and interest.

539 ACKNOWLEDGEMENTS

540 We would like to thank Dr. Kathleen Gallagher for her help filling in for 'Exp2' in some of the
541 tests, in addition to all of the Guide Dogs volunteer puppy walkers, and their dogs, who
542 participated in this study. We would also like to thank the two anonymous reviewers for
543 their valuable feedback on previous versions of this manuscript.

544

545 CONFLICT OF INTEREST STATEMENT

546 The research reported in this publication was funded by Guide Dogs and The University of
547 Nottingham as part of a larger five-year research initiative. Authors of this publication
548 frequently consult with Guide Dogs regarding the behavior of their dogs. Guide Dogs have
549 approved the paper for publication. The terms of this arrangement have been reviewed and
550 approved by the University of Nottingham in accordance with its policies on research.

551

552 AUTHORSHIP STATEMENT

553 NH conceived and designed the study and data collection tools, collected data, performed
554 data analysis and drafted and revised the paper. PC assisted with design of the study,
555 collected data and commented on drafts and revisions of the paper. RS & CM assisted in
556 design of the study, data collection and commented on drafts and revisions of the paper.
557 MG supported statistical analysis and commented on drafts and revisions of the paper. GE
558 initiated the project, monitored the study and commented on drafts and revisions of the
559 paper. LA oversaw the study, conceived and designed the study, monitored data collection,
560 directed data analysis, and drafted and revised the paper.

561 REFERENCES
562
563 Altman, D.G., 1991. Practical statistics for medical research. Chapman and Hall, London
564 Arata, S., Momozawa, Y., Takeuchi, Y., Mori, Y. 2010. Important behavioral traits
565 for predicting guide dog qualification. J. Vet. Med. Sci. 72, 539–545.
566 Arvelius, P., Strandberg, E., Fiske, W.F. 2014. The Swedish Armed Forces temperament test
567 gives information on genetic differences among dogs. J. Vet. Behav.: Clin. Appl. Res.
568 9, 281-289.
569 Asher, L., Blythe, S., Roberts, R., Toothill, L., Craigon, P.J., Evans, K.M., Green, M.J., England,
570 G.C.W., 2013. A standardized behavior test for potential guide dog puppies: Methods
571 and association with subsequent success in guide dog training. J. Vet. Behav.: Clin.
572 Appl. Res. 8, 431-438.
573 Asher, L., Collins, L.M., Ortis-Pelaez, A., Drewe, J.A., Nicol, C.J., Pfeiffer, D.U., 2009. Recent
574 advances in the analysis of behavioral organization and interpretation as indicators
575 of animal welfare. J. R. Soc. Interface 6, 1103-1119.
576 Batt, L.S., Batt, M.S., Baguley, J.A., McGreevy, P.D., 2008. Factors associated with success in
577 guide dog training. J. Vet. Behav.: Clin. Appl. Res. 3, 143-151.
578 Beaudet, R., Chalifoux, A., Dallaire, A., 1994. Predictive value of activity level and behavioral-
579 evaluation on future dominance in puppies. Appl. Anim. Behav. Sci. 40, 273-284.
580 Beerda, B., Schilder, M.B., van Hooff, J.A., de Vries, H., Mol, J.A., 1998. behavioral, saliva
581 cortisol and heart rate responses to different types of stimuli in dogs. Appl. Anim.
582 Behav. Sci. 58, 365-381.
583 Beerda, B., Schilder, M.B., Van Hooff, J.A., de Vries, H.W., 1997. Manifestations of chronic
584 and acute stress in dogs. Appl. Anim. Behav. Sci. 52, 307-319.
585 Bell, A.M., Hankison, S.J., Laskowski, K.L., 2009. The repeatability of behavior: a meta-
586 analysis. Anim. Behav. 77, 771-783.
587 Bryington, A.A., Palmer, D.J., Watkins, M.W., 2002. The estimation of interobserver
588 agreement in behavioral assessment. Behav. Analyst. 3, 323-328.
589 Budaev, S.V. 2010. Using principal components and factor analysis in animal behavior
590 research: Caveats and guidelines. Ethology 116, 472-480.
591 Carter, A.J., Marshall, H.H., Heinsohn, R., Cowlishaw, G., 2012. Evaluating animal
592 personalities: do observer assessments and experimental tests measure the same
593 thing? Behav. Ecol. Sociobiol. 66, 153-160.
594 Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and
595 standardized assessment instruments in psychology. Psychol. Assessment. 6, 284-
596 290.
597 Crone, E. A. 2009. Executive functions in adolescence: inferences from brain and behavior.
598 Developmental Sci., 12, 825-30.
599 Dowling-Guyer, S., Marder, A., D'Arpino, S. 2011. Behavioral traits detected in shelter dogs
600 by a behavior evaluation. Appl. Anim. Behav. Sci. 130, 107-114.
601 Diederich, C. and Giffroy, J. M. 2006. Behavioural testing in dogs: A review of methodology
602 in search for standardisation. App. Anim. Behav. Sci. 97, 51-72.
603 Duffy, D.L., Serpell, J.A., 2012. Predictive validity of a method for evaluating temperament in
604 young guide and service dogs. Appl. Anim. Behav. Sci.138, 99-109.
605 Fratkin, J.L., Sinn, D.L., Patall, E.A., Gosling, S.D., 2013. Personality Consistency in Dogs: A
606 Meta-Analysis. PloS One 8, e54907.

607 Freeman, H., Gosling, S.D., Schapiro, S., 2011. Comparison of Methods for Assessing
608 Personality in Nonhuman Primates. Personality and Temperament in Nonhuman
609 Primates. A. Weiss, J. E. King and L. Murray. New York, Springer, pp. 17-40.
610 Goddard, M.E., Beilharz, R.G., 1984. The Relationship of Fearfulness to, and the Effects of,
611 Sex, Age and Experience on Exploration and Activity in Dogs. Appl. Anim. Behav. Sci.
612 12, 267-278.
613 Goddard, M.E., Beilharz, R.G., 1986. Early prediction of adult behavior in potential guide
614 dogs. Appl. Anim. Behav. Sci. 15, 247-260.
615 Goleman, M., 2010. Use of puppy tests in the evaluation of future dog behavior and
616 character. Med. Weter. 66, 418-420.
617 Gwet, K.L., 2014. Handbook of inter-rater reliability, 4th ed: the definitive guide to measuring
618 the extent of agreement among raters. Advanced Analyitics, LLC, Gaithersburg, MD.
619 Haverbeke, A., De Smet, A., Depiereux, E., Giffroy, J.-M., Diederich, C., 2009. Assessing
620 undesired aggression in military working dogs. Appl. Anim. Behav. Sci. 117, 55-62.
621 Highfill, J., Hanbury, D., Kristiansen, R., Kuczaj, S., Watson, S., 2010. Rating vs. Coding in
622 Animal Personality Research. Zoo Biol. 29, 509-516.
623 Hoffmann, G. 2002. Puppy tests: An evaluation of their predictive validity. PhD, The
624 University of Queensland.
625 Jones, A.C., Gosling, S.D., 2005. Temperament and personality in dogs (Canis familiaris): A
626 review and evaluation of past research. Appl. Anim. Behav. Sci. 95, 1-53.
627 Kim, Y.K., Lee, S.S., Oh, S.I., Kim, J.S., Suh, E.H., Houpt, K.A., Lee, H.C., Lee, H.J., Yeon, S.C.,
628 2010. behavioral reactivity of the Korean native Jindo dog varies with coat colour.
629 Behav. Process. 84, 568-572.
630 King, T., Marston, L.C., Bennett, P.C. 2012. Breeding dogs for beauty and behavior: Why
631 scientists need to do more to develop valid and reliable behavior assessments for
632 dogs kept as companions. pl. Anim. Behav. Sci. 137, 1-12.
633 Marder, A.R., Shabelansky, A., Patronek, G.J., Dowling-Guyer, S., D'Arpino, S. 2013. Food-
634 related aggression in shelter dogs: A comparison of behavior identified by a behavior
635 evaluation in the shelter and owner reports after adoption. Appl. Anim. Behav. Sci.
636 148, 150-156.
637 Martin, P., Bateson, P., 2007. Measuring behavior: an introductory guide, 3rd edition.
638 Cambridge, UK, Cambridge University Press.
639 McCormick, C.M. and Mathews, I.Z. 2010. Adolescent development, hypothalamic-pituitary-
640 adrenal function, and programming of adult learning and memory. Prog. Neuro-
641 Psychoph. 34, 756-765.
642 McCrae, R.R., Costa, P.T., Jr., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M.D., Sanz,
643 J., Sanchez-Bernardos, M.L., Kusdil, M.E., Woodfield, R., Saunders, P.R., Smith, P.B.,
644 2000. Nature over nurture: temperament, personality, and life span development. J.
645 Pers. Soc. Psychol. 78, 173-186.
646 McGarrity, M.E., Sinn, D.L., Gosling, S.D. 2015. Which personality dimensions do puppy tests
647 measure? A systematic procedure for categorizing behavioral assays. Behav. Process.
648 110, 117-124.
649 Miklosi, A. 2011. Dog behavior, evolution, and cognition. New York, Oxford University Press.
650 Mischel, W. 2006. Consistency and Specificity in behavior, in: Funder, D.C., Ozer, D.J. (Eds.),
651 Pieces of the personality puzzle (4th Edition), W.W. Norton & Co., New York, pp. 60-
652 75.

Mornement, K.M., Coleman, G.J., Toukhsati, S., Bennett, P.C. 2014. Development of the behavioral assessment for rehoming K9's (B.A.R.K.) protocol. Appl. Anim. Behav. Sci. 151, 75-83.

Murphy, J. A., 1998. Describing categories of temperament in potential guide dogs for the blind. Appl. Anim. Behav. Sci. 58, 163-178.

Nichols, D.P. 1998. Choosing an intraclass correlation coefficient. Principal Support Statistician and Manager of Statistical Support SPSS Inc. http://www.ats.ucla.edu/stat/Spss/library/whichicc.htm

Overall, K. 2013. Manual of clinical behavioral medicine for dogs and cats. St Louis, Mosby Elsevier Inc., pp. 129-130.

Perez-Guisado, J., Munoz-Serrano, A., Lopez-Rodriguez, R., 2008. Evaluation of the Campbell test and the influence of age, sex, breed, and coat color on puppy behavioral responses. Can. J. Vet. Res. 72, 269-277.

Pfaffenberger, C.J., Scott, J.P., Fuller, J.L., Binsburg, B.E., Bielfelt, S.W., 1976. Guide dogs for the blind: their selection, development and training. Amsterdam, Elsevier.

Scott, J. P. and Fuller, J. L. 1965. Genetics and the Social Behavior of the Dog, Chicago, University of Chicago Press.

Sforzini, E., Michelazzi, M., Spada, E., Ricci, C., Carenzi, C., Milani, S., Luzi, F., Verga, M., 2009. Evaluation of young and adult dogs' reactivity. J. Vet. Behav.: Clin. Appl. Res. 4, 3-10.

Simes, R. J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751-754.

Sinn, D.L., Gosling, S.D., Hilliard, S., 2010. Personality and performance in military working dogs: Reliability and predictive validity of behavioral tests. Appl. Anim. Behav. Sci. 127, 51-65.

Sisk, C.L. and Zehr, J.L. 2005. Pubertal hormones organize the adolescent brain and behavior. Front. Neuroendocrinol. 26, 163-74.

Slabbert, J. M., Odendaal, J. S. J., 1999. Early prediction of adult police dog efficiency - a longitudinal study. Appl. Anim. Behav. Sci. 64, 269-288.

Stamps, J. 2003. behavioral processes affecting development: Tinbergen's fourth question comes of age. Anim. Behav. 66, 1-13.

Stamps, J. & Groothius, G. G. T. 2010. The development of animal personality: relevance, concepts and perspectives. Biol. Rev. 85, 301-325.

Svartberg, K., 2005. A comparison of behavior in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Appl. Anim. Behav. Sci. 91, 103-128.

Svobodova, I., Vapenik, P., Pinc, L., Bartos, L., 2008. Testing German shepherd puppies to assess their chances of certification. Appl. Anim. Behav. Sci. 113, 139-149.

Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav.: Clin. Appl. Res. 1, 94-108.

R. Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Rayment, D.J., De Groef, B., Peters, R.A., Marston, L.C. 2015. Applied personality assessment in domestic dogs: Limitations and caveats. Appl. Anim. Behav. Sci. 163, 1-18.

698 Riemer, S., Muller, C., Viranyi, Z., Huber, L., Range, F. 2014. The predictive value of early
699 behavioral assessments in pet dogs – A longitudinal study from neonates to adults.
700 PloS One 9, e101237.
701 Roberts, C. and McNamee, R., 2005. Assessing the reliability of ordered categorical scales
702 using kappa-type statistics. Stat. Methods Med. Res. 14, 493-514.
703 Uher, J., 2011. Personality in Non-Human Primates: What Can We Learn from Human
704 Personality Psychology? In: Weiss, A., King, J.E., Murray, L. (Eds.), Personality and
705 Temperament in Nonhuman Primates. Atlanta, USA, Springer, pp. 41-76.
706 Valsecchi, P., Barnard, S., Stefanni, C. and Normando, S. 2011. Temperament test for re-
707 homed dogs validated through direct behavioral observation in shelter and home
708 environment. J. Vet. Behav.: Clin. Appl. Res. 6, 161-177.
709 Walter, S.D, Eliasziw, M., Donner, A., 1998. Sample size and optimal designs for reliability
710 studies. Stat. Med. 17, 101-110.
711 Westgarth, C., Pinchbeck, G.L., Bradshaw, J.W., Dawson, S., Gaskell, R.M., Christley, R.M.,
712 2008. Dog-human and dog-dog interactions of 260 dog-owning households in a
713 community in Cheshire. Vet. Rec. 162, 436-442.
714 Wilsson, E. and Sinn, D.L., 2012. Are there differences between behavioral measurement
715 methods? A comparison of the predictive validity of two ratings methods in a
716 working dog program. Appl. Anim. Behav. Sci. 141, 158-172.
717 Wilsson, E. and Sundgren, P.E. 1997b. The use of a behaviour test for the selection of dogs
718 for service and breeding 1: Method of testing and evaluating test results in the adult
719 dog, demands on different kinds of service dogs, sex and breed differences. Appl.
720 Anim. Behav. Sci. 53, 279-295.
721 Wilsson, E. and Sundgren, P. E., 1998a. behavior test for eight-week old puppies -
722 heritabilities of tested behavior traits and its correspondence to later behavior.
723 Appl. Anim. Behav. Sci. 58, 151-162.
724 Wilsson, E. and Sundgren, P. E., 1998b. Effects of weight, litter size and parity of mother on
725 the behavior of the puppy and the adult dog. Appl. Anim. Behav. Sci. 56, 245-254.

726 **Table 1** Description of the juvenile guide dog behavior test protocol**.** Subtests 1-7 were conducted indoors in the first test arena, and subtests 8-11 were conducted in the second test arena.

Subtest	Description
Dog and PW enter the room (test arena) with the dog on the lead and the test begins	
1) Meet a stranger *(on lead)*	PW and dog entered test area, and approached Exp1 who stands at the opposite end of the arena to the entry door. Both PW and dog invited to greet Exp1. Dog greeted by: holding out hand (under head); making brief eye contact; smiling and petting dog calmly. While explaining the test process to the PW dogs were softly petted on head, only if they approached.
2) Obedience– PW *(on lead)*	PWs were instructed to walk the dog, on the lead, around the test arena and to ask for: a 'sit'; a 'sit-wait'; and a 'down', at marked stations.
3) Obedience – STR *(on lead)*	Exp1 took dog and repeated obedience commands from subtest 2. Hand signals were used in conjunction with the commands wait* (palm up), and down (point down for requests 1 and 2, place pointed hand on floor for request 3). Commands repeated a maximum of three times before using treats; with 5s intervals between repeated commands.
4) Path *(on lead)*	Exp1 led the dog towards foam path, with the dog lined up to walk over and across the foam path. Lead tension was loose so the dog could avoid or step off the path if it chose to. The procedure was repeated twice for all dogs. A hand lure was used where the dog's attention was elsewhere or their actions (i.e. jumping) appeared to place it at risk of tripping over the path edges. If the dog actively avoided the path, or got off less than two thirds across, a hand or treat lure was used and procedure was repeated up to three times.
Dog is given a 2-3 minute break and offered a bowl of water. Following which two play behavior subtests occurred (data not presented here))	
5) Body check *(off lead)*	Exp1 and Exp2 knelt down and called dog to them. Exp2 held the dog's collar and/or used a treat lure to keep dog still (where required) while Exp1 conducted the physical examination which included: a slow pet to the head; ears were then smoothed and lifted for inspection; the dog was then stroked down its back, sides, chest then legs where paws were lifted and given a slight press (attempted twice only). Exp1 & Exp2 avoided eye contact with the dog, talked soothingly and if unable to conduct the subtest waited up to one minute for the dog to calm down before carrying on.
6) Head ring *(off lead)*	Exp1 called dog to them and placed the ring c.20cm in front of the dogs face. Exp1 inserted hand through ring to place treat in front of dogs' muzzle, at which point hand was slowly pulled back through the ring and stopped when dogs head was (or could be) fully inserted. Repeated twice.
7) Tea-towel *(off lead)*	Exp1 called dog to them and offered it a treat and, While dog retrieved the treat, Exp1 placed a quarter folded tea-towel over its back. Exp1 remained in position for 10s or until the dog removed the tea-towel. Repeated twice.
Dog is given a 2-3 minute break and offered a bowl of water. Following which dog is put back on the lead and led to the second testing arena to the beginning of subtest 8 by Exp1	

Subtest	Description
8) Food *(on lead)*	Exp1 led dog to the cones and asked to "sit" once. Exp1 and dog stayed there for 10s then dog led forward and walked past the plates. If the dog stopped or tried to reach the plates Exp1 stopped ahead of it, holding it back from the food, turned and calls dogs name, if no response then the following commands were used; "come on", followed with "dogs name" and " leave". If the dog refused to leave the plates it is touched on the side flank to gain its attention then finally lured away with a treat (only if required).
9) Robin *(on lead)*	As Exp1 and the dog approached within 0.3m of the stimuli Exp3 activated the remote control pulling device. The toy robin emerged from a hide to the right and rapidly moved across dogs' path to hide again on the dogs left. If the dog stopped or tried to reach the robin the response procedure from subtest 8 was repeated until the dog moved on.
10) Pigeons *(on lead)*	Exp1 and dog walked past two plastic pigeons placed 0.5 meters from the path. If the dog stopped or tried to reach them the procedure from subtest 8 was repeated until the dog moved on.
11) Human distraction *(on lead)*	Exp1 & dog walked past Exp3 who stood 1/2 a meter from the path. Exp3 stood still and looked at the dog as they approach but withholds any other contact. If the dog stopped or jumped up on Exp3 the response procedure from subtest 8 was repeated until the dog moved on.

Note: PW indicates the dog's puppy walker, Exp indicates an experimenter, STR is used to represent Exp1, in subtest 1 and 3, who was previously unknown to the dogs and therefore acts as a stranger (STR) for subtests 1 and 3. *dog asked to "sit", once sat dog asked to "wait", Exp1 then takes two steps away from the dog, holding a long lead, repeats "wait" then returns to dog and praises. If, at any point, the dog jumped up onto the experimenters they would turn their back on the dog, cross their arms and wait until jumping ceased then resume the test calmly. Between subtest 4 and 5 the dogs took part in two further subtests on play behavior carried out by Exp2, the results of which do not form part of this study.

727

728

729

730

731

732 **Table 2** A list of behavioral measures that comprised the juvenile guide dog behavior test ethogram. Still frame images of the postural measures can be found in online supplementary material
733 (supplementary figures 1-16)

Subtest	Behavior/Measure	Type	Definition
All	Jumps	Continuous (count)	Dog's front two, or more, paws off the ground simultaneously (but not when rearing due to strong lead pulling)
	Whines	Continuous (count)	Frequency of whining bouts, with a bout defined as: a continuous emission of whining ending when whining stops
	Scratches	Continuous (count)	Dog scratches itself with back feet
	Barks	Binary	Scored as whether a bark was observed at any point during the video scoring
	Shakes	Binary	Shakes head or whole body (in subtests 5-7 only)
Subtest 1- Meet a Stranger	Pull Strength (greet)	Categorical	None - Lead relaxed, dog not straining against the lead or collar; Slight - Head extended forwards & lead tense but weight evenly distributed over all four feet; Extreme- weight forwards over front legs, rear legs pushing and/or one or more front paws raised off the floor.
	Low Posture	Binary (1/0)	Low posture during greeting: front legs bent; tail neutral or low AND wagging; head lowered and ears backwards
Subtest 2 and 3 - Obedience	Sit/Wait/Down Response	Categorical	Dog obeyed 'sit' 'wait' or 'down' command and sits on hind quarters in response to (1) first command; (2) second command or more; (3) does not respond to command appropriately
	Gaze Proportion	Continuous (%)	The proportion of time spent gazing at the face of the handler, relative to the total length of the subtest
	Lip-licks	Frequency	Tongue briefly seen outside of mouth, sweeping across lips/muzzle or up to nose
Subtest 4 - Path	Crossed	Binary (1/0)	Dog walked on the path from one end to the other
Subtest 5- Body Check	Score	0-6	Number of body parts out of a maximum of six successfully checked
	Mouths	Continuous (count)	Low pressure, non-injurious grab of testers limbs or clothes with mouth. Recorded as a count of total number observed
	Licks	Binary (1/0)	Licking of Experimenters limbs or clothes. Recorded as 0/1 for each of the six body parts checked
Subtest 6 - Head Ring	Ear Position	Binary (1/0)	Neutral - individual relaxed ear state, neither forwards nor backwards facing; Backwards - ears flattened backwards against the head, exposing the inner ear lining to view
	Tail Height	Categorical	Neutral - relaxed tail allowed to fall vertically from where the tail joins the spine; Half Up - he tail falling below the level of the dog's back, but raised from neutral; Up - tail in line with, or above, the level of the dog's back
	Body Posture	Binary (1/0)	Neutral- weight evenly distributed, head not extended; or Stretched- weight over front legs and head extended, when head inserted in ring
Subtest 7 – Tea-Towel	Attempts to Remove	Binary (1/0)	An attempt by the dog, successful or not, to remove the tea-towel from their back
	Turns Head	Binary (1/0)	Head turned to look at the tea-towel with no attempt to remove
	Change from Neutral	Binary (1/0)	Dog's body posture changed from neutral when tea-towel placed on back. Changes included: arched back; lowered tail and backwards ears
	Plays with	Binary (1/0)	Dog played with the tea-towel after removal. Play included: shaking; tearing at or running with the tea-towel held in mouth
Subtest 8 - Subtest 10: Food; Robin & Pigeons Distractions	Pull Strength (distraction)	Categorical	None – lead may be tense but dog's weight evenly distributed across all four feet and no straining against the lead; Medium – head extended towards stimulus, weight pushing forwards and straining against the lead, all paws remain on the ground; Strong –weight forwards, the dog is straining against the lead with head extended towards stimulus, back legs are stretched and one or more front paws raised off the floor
	Time Oriented	Continuous (seconds)	Time the dog remained oriented towards the stimulus, with head or head & body, after first recall prompt
	Approaches	Binary (1/0)	Dog left side of Experimenter and walked towards the stimulus
	Avoids	Binary (1/0)	Dog actively avoided the stimulus by backing away or walking closer to Experimenter (not observed in subtest 8)
Subtest 11: Human Distraction	Pull Strength (greet)	Categorical	As above
	Jumps	Binary (1/0)	Jumped up with front paws placed on human distraction (Exp 3)

Table 3 Results of Cohens Kappa (*K*) analysis for inter-rater reliability for the 16 binary variables of the juvenile guide dog behavior test (n=40 test videos). For subtests 5, 6 & 7 measures were repeated within the subtest so were combined, resulting in larger degrees of freedom for these measures.

Subtest	Variable	*K*	SE	df	*p*	Lower 95% CI	Upper 95% CI
All	Barks	0.62	0.17	39	<0.001	0.29	0.94
1: Meet a stranger	Low posture	0.77	0.15	39	<0.001	0.47	1.07
2: PW Obedience	"Sit" performance	0.70	0.13	39	<0.001	0.45	0.94
	"Wait" performance	0.68	0.17	37	<0.001	0.35	1.02
4: Path	Crossed	0.90	0.10	39	<0.001	0.69	1.10
5: Body Check	Licks	0.35	0.16	235	<0.001	0.05	0.66
	Mouths	0.71	0.08	235	<0.001	0.54	0.87
6: Head Ring	Body posture	0.54	0.15	77	<0.001	0.24	0.84
	Ear position	0.85	0.06	76	<0.001	0.73	0.96
7: Tea-towel	Attempts to remove	0.84	0.63	79	<0.001	-0.40	2.07
	Change from neutral	0.62	0.13	79	<0.001	0.36	0.87
	Plays with	0.80	0.07	79	<0.001	0.67	0.93
	Turns	0.62	0.09	79	<0.001	0.45	0.79
10: Pigeons	Approaches	0.63	0.20	39	<0.001	0.24	1.01
11: Human	Jumps	0.94	0.06	39	<0.001	0.84	1.05
5-7: Body sensitivity	Shakes	1.00	0.00	39	<0.001	1.00	1.00

Table 4 ICC coefficients, degrees of freedom, confidence intervals and confidence interval width for intra-rater reliability assessment of all continuous and ordinal variables from the of the juvenile guide dog behavior test (n=40 test videos). Continuous variable were assessed using the consistency method and single measure ICC values are reported, while for ordinal variables absolute agreement was applied and average measures are reported to achieve a mean weighted kappa.

Subtest	Variable	Data Type	ICC	df	Lower 95% CI	Upper 95% CI	95% CI Width
All	Jumps	Continuous	0.97	39	0.94	0.98	0.05
	Mouths	Continuous	0.88	39	0.78	0.93	0.15
	Scratches	Continuous	0.72	39	0.53	0.84	0.31
	Whines	Continuous	0.93	39	0.87	0.96	0.10
1: Meet a stranger	Pull strength	Ordinal	0.87	38	0.73	0.93	0.20
2: PW Obedience	"Down" performance	Ordinal	0.92	39	0.84	0.96	0.11
3: STR Obedience	"Sit" performance	Ordinal	0.93	39	0.86	0.96	0.10
	"Wait" performance	Ordinal	1.00	39	-	-	-
	"Down" performance	Ordinal	0.97	39	0.95	0.96	0.01
6: Head Ring	Tail height	Ordinal	0.63	77	0.42	0.76	0.35
8: Food	Pull strength	Ordinal	0.88	39	0.77	0.94	0.17
	Time oriented	Continuous	0.98	39	0.97	0.99	0.02
9: Robin	Pull strength	Ordinal	0.79	38	0.60	0.89	0.29
	Time oriented	Continuous	0.99	38	0.97	0.99	0.02
10: Pigeons	Pull strength	Ordinal	0.63	39	0.31	0.80	0.49
	Time oriented	Continuous	0.95	39	0.90	0.97	0.07
11: Human	Pull strength	Ordinal	0.73	39	0.48	0.86	0.38

Table 5 Test-retest correlations from the five and eight month juvenile guide dog behavior tests that achieved significance after Improved Bonferroni correction. *r* indicates the correlation coefficient; *P*, p values and *cP* corrected p-values using the Improved Bonferroni procedure; NS indicates Not Significant.

Subtest	**Behavior/Measure**	**Test**	***r***	***P***	***cP***
All	Jumps	Spearman's	0.67	<0.001	<0.001
	Barks	Kendall's tau-b	0.46	<0.001	<0.001
	Whines	Spearman's	0.29	0.017	NS
Subtest 1- Meet a Stranger	Low posture	Kendall's tau-b	0.49	<0.001	<0.001
Subtest 5- Body Check	Mouths	Kendall's tau-b	0.33	0.005	NS
	Licks	Kendall's tau-b	0.35	0.005	NS
Subtest 6 - Head Ring	1st Ear position	Kendall's tau-b	0.45	<0.001	0.002
	1st Tail height	Kendall's tau-b	0.41	0.001	0.012
	1st Body posture	Kendall's tau-b	0.26	0.032	NS
	2nd Ear position	Kendall's tau-b	0.37	0.002	0.034
	2nd Tail height	Kendall's tau-b	0.38	0.002	0.030
Subtest 7 - Tea towel	1st Attempts to remove	Kendall's tau-b	0.44	<0.001	0.003
	1st Change from neutral	Kendall's tau-b	0.36	0.003	NS
	1st Plays with	Kendall's tau-b	0.46	<0.001	0.001
	2nd Attempts to remove	Kendall's tau-b	0.60	<0.001	<0.001
	2nd Change from neutral	Kendall's tau-b	0.49	<0.001	0.001
	2nd Plays with	Kendall's tau-b	0.39	0.002	0.030
Subtest 8 - Food	Pull strength	Kendall's tau-b	0.34	0.002	0.021
	Time oriented	Spearman's	0.50	<0.001	<0.001
	Approaches	Kendall's tau-b	0.27	0.027	NS
Subtest 9 - Robin	Pull strength	Kendall's tau-b	0.34	0.003	0.046
	Time oriented	Spearman's	0.32	0.010	NS
	Approaches	Kendall's tau-b	0.61	<0.001	<0.001
Subtest 10 - Pigeons	Pull strength	Kendall's tau-b	0.36	0.001	0.011
Subtest 11- Human distraction	Jumps	Kendall's tau-b	0.38	0.002	0.028

Table 6 Results of binary logistic regression models; predictors significant to p<0.1. Dependant variable was withdrawal for behavior vs. qualification (n=52Q, 21W-B).

Age (m)	Subtest	Measure	Wald	*P*	OR	95% CI	% Odds Change	Unit
5	All	Barks	3.86	0.049	3.76	(1.00,14.08)	276%	Binary
	2: PW Obedience	Lip-licks	8.54	0.004	1.62	(1.17, 2.25)	62%	Per lick
	2: PW Obedience	Gaze proportion	4.1	0.043	0.96	(0.92, 1.00)	-4%	Per %
	2: PW Obedience	'Down' performance (1^{st} vs. none)	3.09	0.079	0.38	(0.13, 1.12)	-62%	Categorical
	3: STR Obedience	'Down' performance (1^{st} vs. 3^{rd})	3.14	0.076	0.31	(0.8, 1.13)	-69%	Categorical
	5-7: Body sensitivity tests	Shakes	4.85	0.028	0.25	(0.87, 1.14)	-75%	Binary
	7: Tea-towel	2^{nd} Attempts to remove	4.14	0.042	2.98	(1.04, 8.52)	198%	Binary
	7: Tea-towel	2^{nd} Plays with	4.16	0.046	2.92	(1.02, 8.30)	192%	Binary
	7: Tea-towel	1^{st} Attempts to remove	3.48	0.062	2.96	(0.95, 9.28)	192%	Binary
	8: Food	Time oriented	4.15	0.045	1.04	(1.00, 1.09)	4%	Per second
8	1: Meet a Stranger	Low Posture	4.78	0.028	0.16	(1.21, 0.83)	-84%	Binary
	7: Tea-towel	2^{nd} Turns	4.04	0.039	0.27	(1.07, 0.94)	-73%	Binary
	7: Tea-towel	1^{st} Change from neutral	3.60	0.058	6.23	(0.94, 41.20)	623%	Binary
	7: Tea-towel	1^{st} Plays with	3.04	0.081	2.78	(0.88, 8.75)	178%	Binary
	8: Food	Time oriented	4.77	0.027	1.05	(1.01, 1.10)	5%	Per second
	8: Food	Pull (Slight vs. None)	4.74	0.029	6.35	(1.20, 33.55)	535%	Categorical
	9: Robin	Approach	3.60	0.058	0.16	(0.02, 1.06)	-84%	Binary
	9: Robin	Pull (Slight vs. Strong)	4.08	0.043	0.23	(0.05, 0.96)	-77%	Categorical
	10: Pigeons	Avoid	4.89	0.022	5.89	(1.32, 29.78)	489%	Binary
	10: Pigeons	Pull (Strong vs. Slight)	3.10	0.078	0.24	(0.50, 1.17)	-76%	Categorical

Table 7 Rotated component matrix showing item loadings from a principal components analysis on six correlated variables from the 5 month juvenile guide dog behavior test. All variables were significantly associated with other and were associated with Guide Dogs' training outcome (qualification or withdrawal for behavior) to a significance of p<0.1.

		Component	
Subtest	**Variable**	**5M-PC1**	**5M-PC2**
7: Tea-towel	2nd Attempts to remove	**0.868**	0.141
7: Tea-towel	2nd Plays with	**0.864**	-0.009
7: Tea-towel	1st Attempts to remove	**0.783**	0.028
All	Barks	0.112	**0.689**
2: PW obedience	Lip-licks	0.242	**0.653**
5-7: Body sensitivity tests	Shakes	0.264	**-0.626**

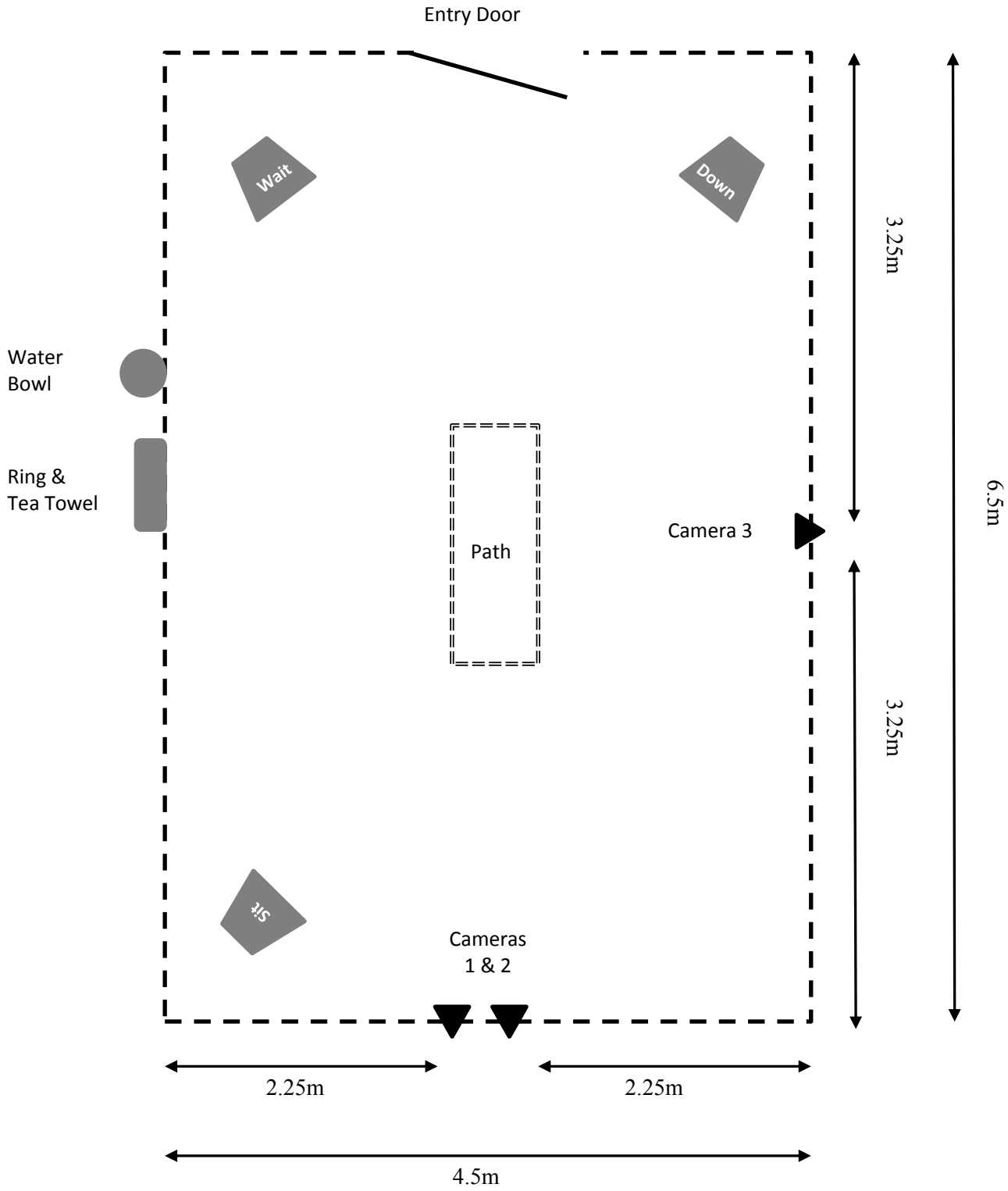
Table 8 Rotated component matrix showing item loadings from a principal components analysis on nine correlated variables from the 8 month juvenile guide dog behavior test. All variables were significantly associated with other and were associated with Guide Dogs' training outcome (qualification or withdrawal for behavior) to a significance of p<0.1.

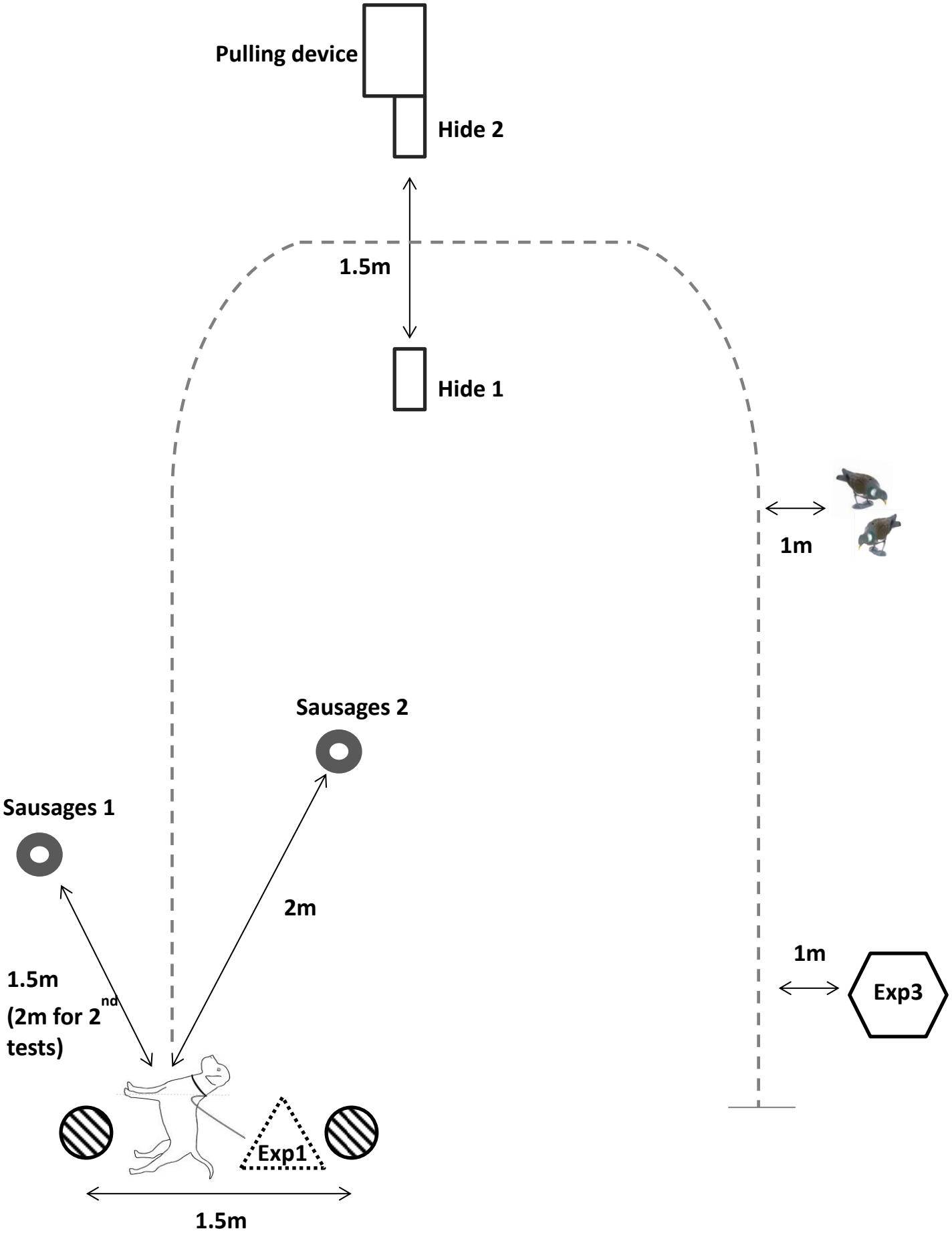
		Component		
Subtest	**Variable**	**8M-PC1**	**8M-PC2**	**8M-PC3**
9: Robin	Pull strength	**0.868**	-0.062	-0.038
7: Tea-towel	1st Plays with	**0.746**	0.260	-0.125
8: Food	Pull strength	**0.695**	-0.260	-0.211
10: Pigeons	Pull strength	**0.609**	-0.350	-0.011
10: Pigeons	Avoids	-0.205	**0.780**	-0.065
7: Tea-towel	1st Change from neutral	0.040	**0.720**	0.110
7: Tea-towel	2nd Turns	-0.212	0.065	**0.812**
9: Robin	Approach	0.472	-0.339	**0.577**
8: Food	Time oriented	0.431	-0.239	-0.474

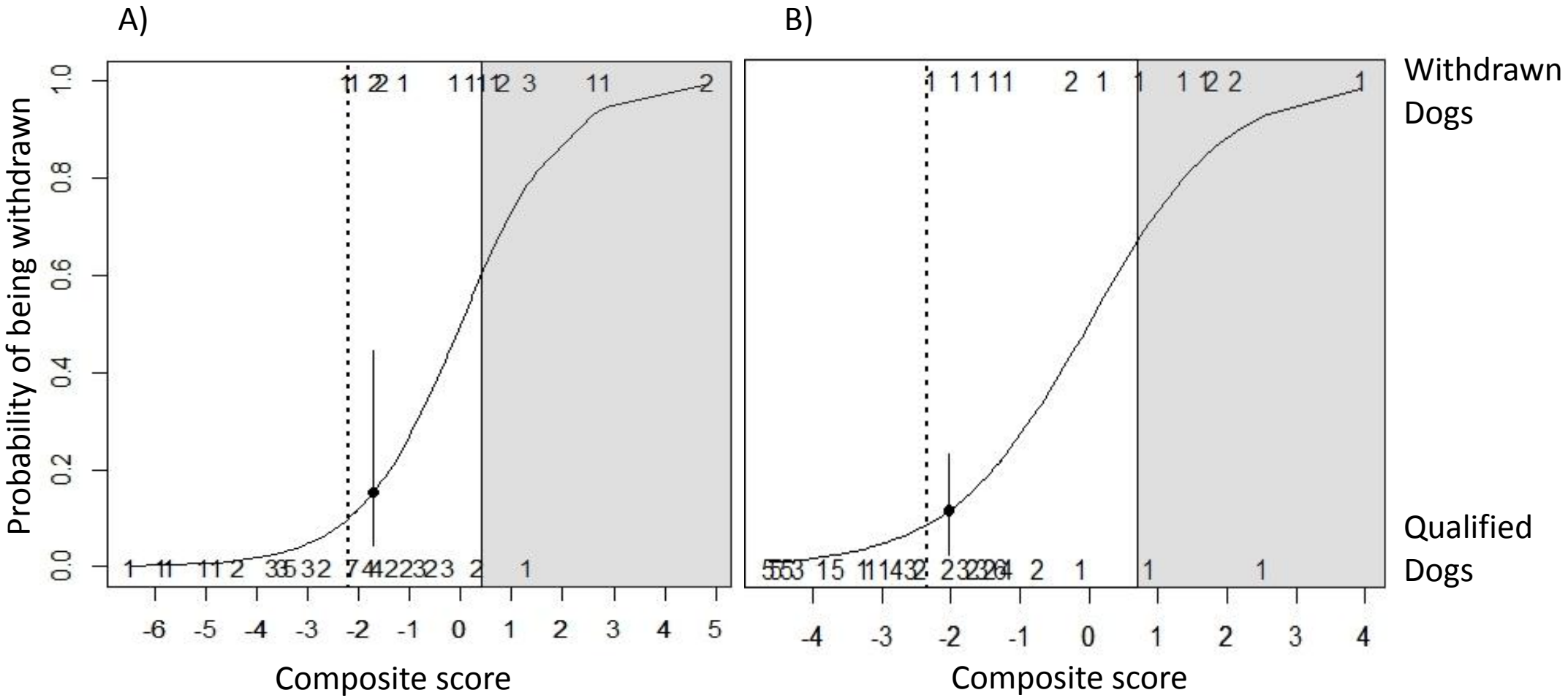
Figure 1 Schematic representation of experimental set up for the first test arena, shown from above. Chairs were used to mark the outer perimeter and signs placed upon chairs marked the locations for the obedience commands. The path was placed in the centre only for that test, for all other tests it was placed upon chairs next to the other equipment. The position of the entry door represents its position in the majority of venues. Subtests 1-7 were conducted in this test arena.

Figure 2 Schematic representation of experimental set up of the second test arena for the final four subtests (8-11), shown from above. The dashed circles represent two small cones that mark the beginning of the subtest 8. The distances given remained constant whilst the distances between stimuli varied according to space available. 'Sausages 2' were removed for the second test, due to the dogs increased size and strength, and 'Sausages 1' was moved to 2m from the dogs start position. The dotted line represents the path taken through the subtests. The triangle indicates the position of Exp1 relative to the dog throughout the subtests.

Figure 3 Probability of a dog qualifying in guide dog training or being withdrawn for behavioral reasons plotted against the dogs actual training outcome and: A) Composite score from logistic regression in 5 month old dogs (including: time oriented towards the food (subtest 8), down performance (subtest 2 & 3); and two component scores from PCA (5M-PC1 and 5M-PC2)); B) Composite score from logistic regression in 8 month old dogs (including: three component scores, 8M-PC1, 8M-PC2, 8M-PC3, and one independent variable; 'low' greeting posture from subtest 1). The numbers inside the plots represent individual dogs placed according to their composite score. The dotted lines indicate a 10% probability point; dogs to the left of which have a less than 10% chance of being withdrawn. The grey boxes include all dogs with a chance of being withdrawn for behavior, greater than: A) 60% and B) 50%. Such thresholds are given as an example of how these models could be utilised to aid decision making within Guide Dogs by alerting those dogs with the highest and lowest chances of being withdrawn.


Entry Door
Wait
Down
3.25m
6.5m
Water
Bowl
Ring &
Tea Towel
Path
Camera 3
3.25m
Sit
Cameras
1 & 2
2.25m
2.25m
4.5m


Pulling device
Hide 2
1.5m
Hide 1
1m
Sausages 2
Sausages 1
2m
1.5m
(2m for 2nd
tests)
1m
Exp3
Exp1
1.5m

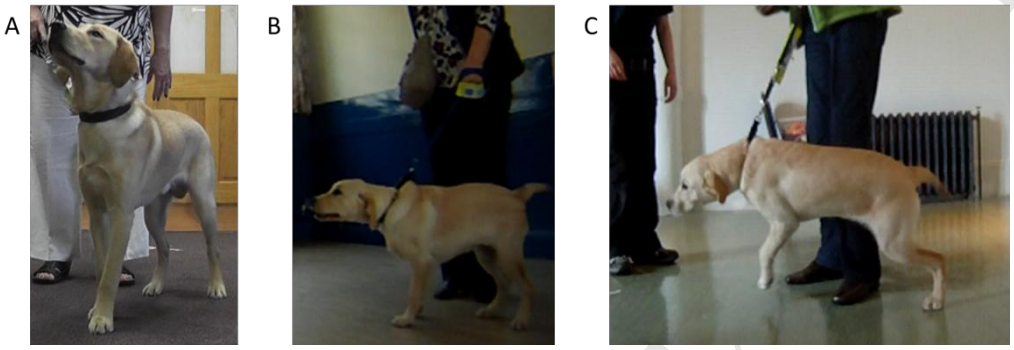

A)
B)
Probability of being withdrawn
0.0
0.2
0.4
0.6
0.8
1.0
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
Composite score
Composite score
Withdrawn Dogs
Qualified Dogs

HIGHLIGHTS

- A practical behavior test for juvenile guide dogs is described and evaluated
- Behaviors that demonstrate consistency across time are highlighted
- Dogs likely to qualify or be withdrawn from training were successfully identified
- The test is a reliable and valid method for testing the behavior of juvenile dogs

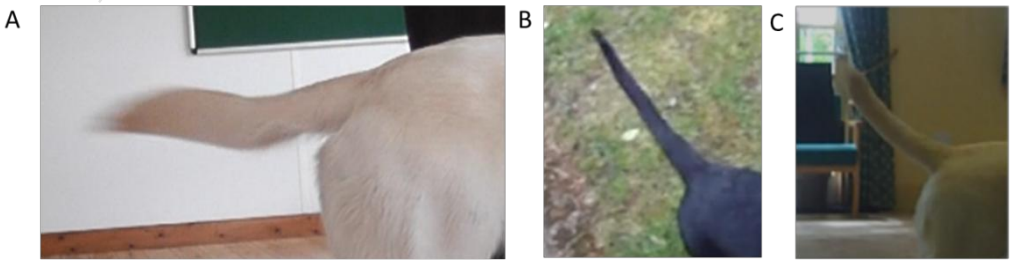
Supplementary material – Ethogram examples

This document includes still frame images taken from video footage of the juvenile guide dog behaviour test. These images are provided as a supplement to the ethogram described in Table 2 of the manuscript.

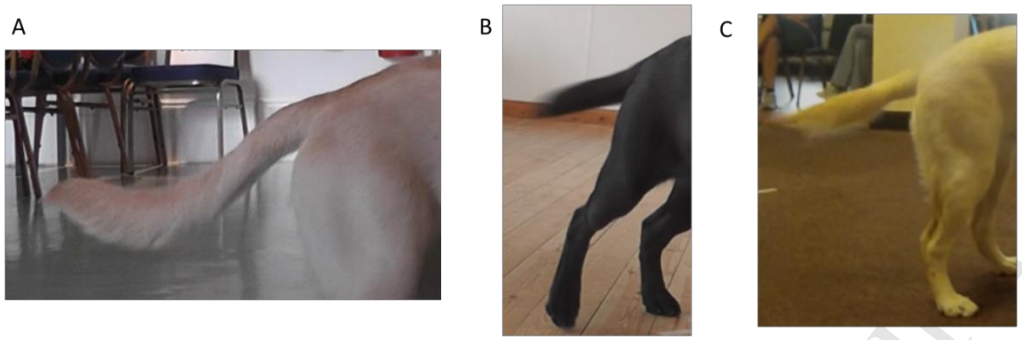


Supplementary Figure 1 Still frame examples of the three categories of pull strength observed during subtest 1 ('Meet a Stranger'). A, shows no pull; B, is showing slight pull; and C, is pulling strongly.

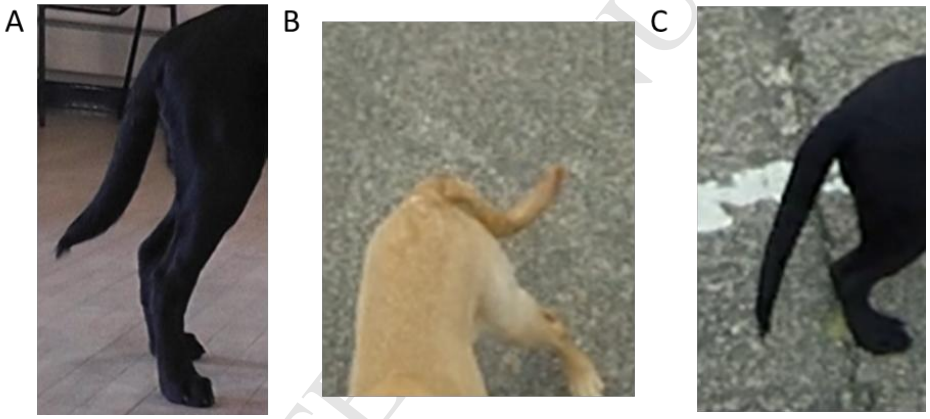
Supplementary Figure 2 Still frame example of a 'Low' posture observed during greeting in subtest 1 ('Meet a Stranger'). The dogs: front legs were bent; tail neutral or low and wagging; head lowered and ears backwards.



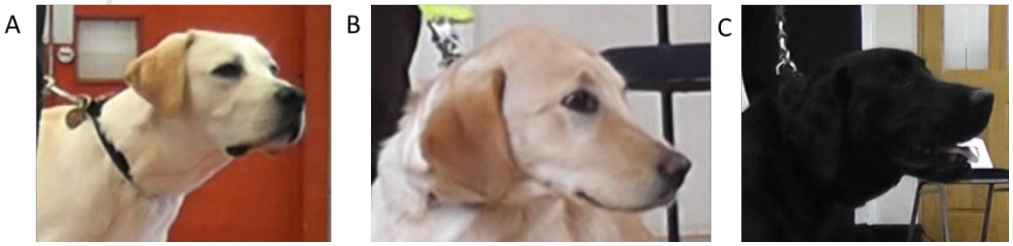
Supplementary Figure 3 Three still video image examples of a tail classified as being 'Up'. This was defined as the tail being in line with, or above, the level of the dogs back.



Supplementary Figure 4 Three still video image examples of a 'Half Up' tail, defined as: the tail falling below the level of the dogs back, but raised from neutral.

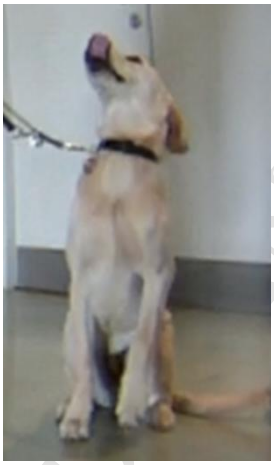


Supplementary Figure 5 Three still video image examples of 'Neutral' tails, which was defined as being a relaxed tail falling vertically from where the tail joins the spine. A low tail was a defined as a tail that was curled under/in between the dogs' legs; however low tails were not observed during any of the juvenile guide dog behaviour tests.



Supplementary Figure 6 Three still video image examples of 'Neutral' ear positions. (A) represents a borderline ear position, in this case the dog in question had smaller ears and when investigated further this dog retained a slightly forwards positions throughout testing from which point they would move forward or backward so this position was taken as 'neutral' for this dog. (B) A standard neutral ear position. (C) Note the comparative difficulty of ear visibility on black coated dogs.

Supplementary Figure 7 Three still video image examples of 'Backwards' ear positions.



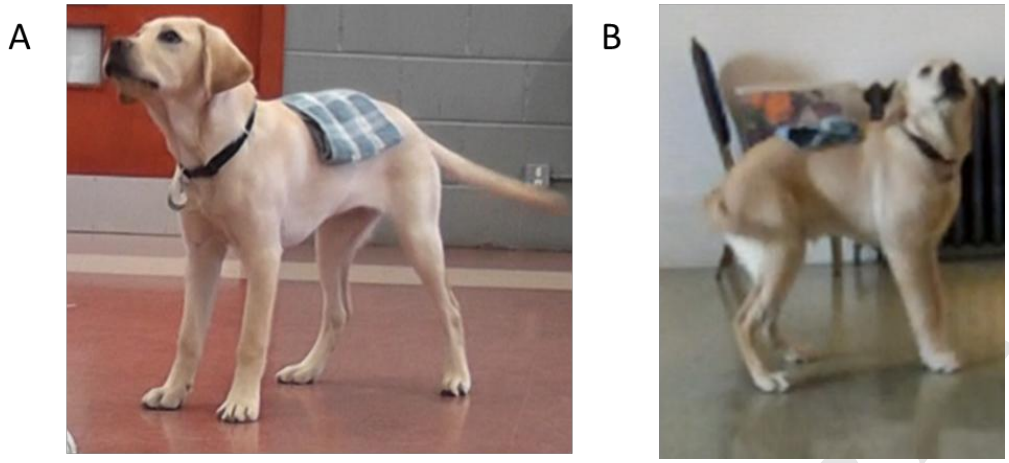
Supplementary Figure 8 Still video image of a dog showing a lip-lick during the PW obedience section (subtest 2).

Supplementary Figure 9 Still video image of a gaze towards the puppy walker (PW) during the 'Down' command during subtest 2, the dogs PW is located outside the image to the left.

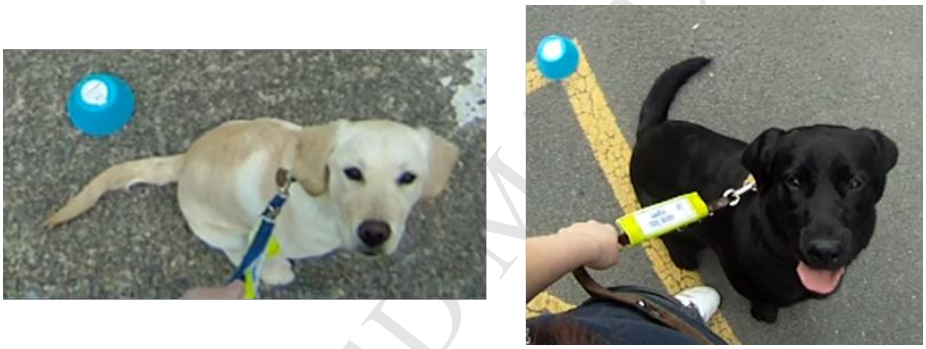
Supplementary Figure 10 Still video image of a gaze towards the experimenter (Exp1) during the 'Wait' command in subtest 3; Exp1 is located outside the image to the right holding the lead.

Supplementary Figure 11 Still video image example of a dog showing full head insertion with body 'Stretched' when offered a treat through a ring during subtest 6 'Head Ring'.

Supplementary Figure 12 Still video image example of a dog showing full head insertion with body 'Neutral' when offered a treat through a ring during subtest 6 'Head Ring': ears 'Backwards'; tail 'Up'; body 'Neutral' with weight evenly distributed.

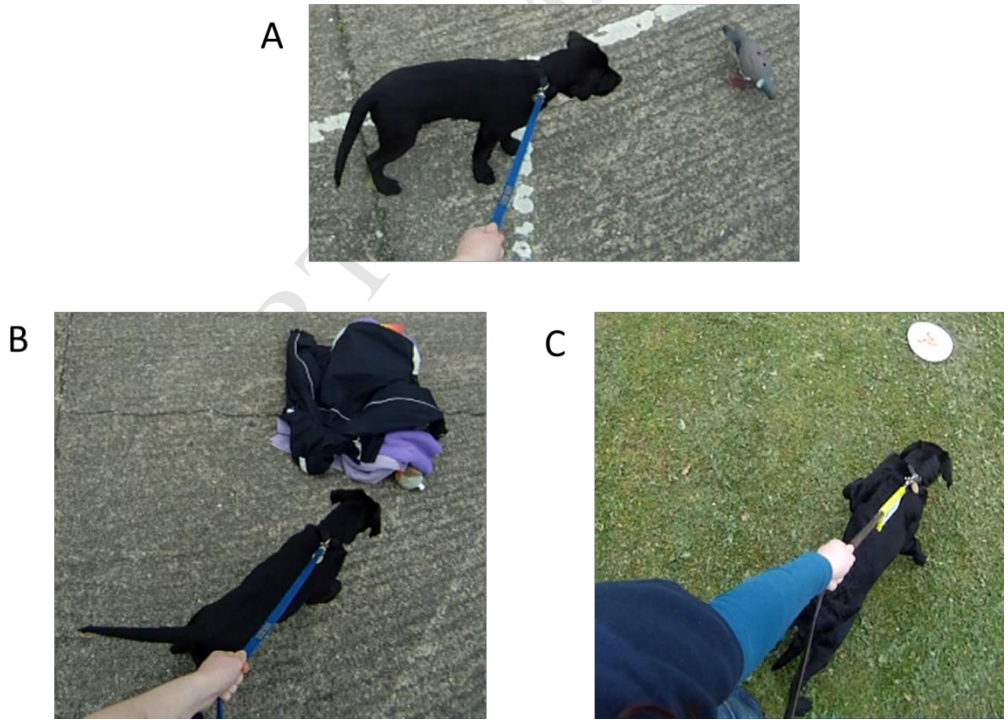


Supplementary Figure 13 Still video image of the two body postures recorded during the tea-towel subtest (subtest 7): (A) shows a dog with a neutral posture unchanged since before application of the tea-towel to the dogs back; (B) shows a dog with a posture that changed from neutral upon application of the tea-towel: ears are backward, tail low and back arched.



Supplementary Figure 14 Still video images from the head-cam worn by Exp1 showing two examples of gazing towards Exp1 during subtests 8-11.

Supplementary Figure 15 Still video images from the head-cam worn by Exp1 showing a dog that had stopped walking and oriented towards the 'pigeons' (subtest 10).



Supplementary Figure 16 Still video images from the head-cam worn by Exp1 showing the three different strengths of pull categorised in subtests 8-11. A) Shows a dog categorised as not pulling; weight is evenly distributed across all four feet. B) Shows a 'Medium' pull strength; weight is pushing forwards and dog is attempting to reach the stimulus. C) Shows a strong pull; weight forwards, lead tense, back legs stretched and pushing with both front paws off the ground.