The Thesis Committee for Stephen Nicholas Debono
Certifies that this is the approved version of the following thesis:

Predicting Success in a Detector-Dog Program: Subjective Ratings of Puppies and Characteristics of Handlers

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Samuel D. Gosling

David L. Sinn

Predicting Success in a Detector-Dog Program: Subjective Ratings of Puppies and Characteristics of Handlers

by

Stephen Nicholas Debono, B.A.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May, 2014

Dedication

To Moatie, whose adoption in 2002 started me on this path, and who has joyfully stuck with me for every twist and turn, no matter what it meant to her life. As dogs do.

Acknowledgements

There are just so many individuals to thank. This list is just the tip of the iceberg.

First, I have to express a massive thank you to Sam Gosling for giving me, a dog trainer with what some might call a very "non-traditional" academic background, this enormous opportunity. Sometimes people get lucky, and I consider myself very lucky to have somehow shown up on your radar.

Second, I owe David Sinn a ton of thanks for all his persistent guidance, honesty, and advice.

Third, thank you to everyone at Australia Customs and Border Protection Service, especially Michelle and Smyl Fischer, for making the research contained in this thesis possible.

Fourth, I thank my parents, Sam and Maria Debono, for their loving support throughout the years. I definitely couldn't have made it here without you.

Fifth, I want to send a huge thank you out to all my friends, who have been so supportive throughout my semi-regular bouts of self-doubt. Miles Bensky, especially, talked me down from more cliffs than I can count. And of course, my wonderful girlfriend Robin Cupell, for having the patient and forgiving heart to date me.

Sixth, I want to acknowledge all my wonderful pets, both present: Moatie, Sonny, Alan, Fern, and past: Bucket, Banana, Charlie M., for having brought me happiness and cheer whenever you were around. And for being really really cute.

Lastly, I am so grateful to everyone from my parallel dog-training universe. Suzanne Clothier, you've been extraordinarily influential to my work. To my clients, both dog and human, I thank you for showing extraordinary amounts of trust in me, even when I make mistakes. My goal is to never take that trust for granted.

Abstract

Predicting Success in a Detector-Dog Program: Subjective Ratings of Puppies and Characteristics of Handlers

Stephen Nicholas Debono, M.A.

The University of Texas at Austin, 2014

Supervisor: Samuel D. Gosling

Detector-dog organizations continually work to improve their effectiveness. Detector dogs commonly work in partnership with human handlers. Organizations spend considerable amounts of resources selecting both dogs and humans suited for the required duties. This thesis describes two studies. In the first study, we developed and evaluated a subjective dog trait-rating survey to obtain ratings of dogs by the people raising them. In the second study, we examine how human characteristics relate to job performance for professional detector-dog handlers.

In working-dog breeding programs, candidate puppies are often placed with volunteer families (puppy raisers) who care for and raise the puppies. These families have extensive opportunities to observe a puppy's behavior across time so they may be able to make accurate trait evaluations, which could predict subsequent performance. In Study 1, we develop, implement, and evaluate the Puppy Raiser Subjective Survey (PRS Survey) on a population of puppy raisers from a large detector-dog organization (Australian Customs & Border Protection Service; AC&BPS). Analyses identified seven dimensions

of personality but a model including these variables was not able to significantly predict working performance.

Selecting people who are suited to work as dog handlers is likely to be important to the success of working-dog programs. Detector-dog programs often undergo a resource intensive process to select the best humans for the job. However, there has been scarce research on the types of traits that make one handler more effective than another. In Study 2, we develop, implement, and evaluate an instrument used to identify human characteristics that predict success as AC&BPS detector-dog handlers. We show that job seniority was the strongest predictor of detector-dog handler job performance. We also show intriguing possibilities that participation in a greater number of sports, particularly at competition levels, may correlate with better job performance.

Table of Contents

List of Tables

List of Figures

Chapter 1

General Introduction

Dogs' acute olfactory senses make them highly effective at identifying individual scents, even when the scents are masked by several other odors (Furton et al., 2002; Harper, Almirall, & Furton, 2005; Helton, 2009). As a result, *odor-detector dogs* are deployed by numerous military organizations and police forces (Harper et al., 2005; Helton, 2009) to find illicit contraband such as narcotics and explosives. It is well established that some dogs are more behaviorally suited for odor-detector dog work than are others (Furton & Myers, 2001; Jones & Gosling, 2005). In many detector dog and police organizations, over half of candidate dogs that undertake training are disqualified due to behavior deemed unsuitable for the required tasks (Maejima et al., 2007; Slabbert & Odendaal, 1999; Wilsson & Sundgren, 1997). Raising and training dogs consumes a large amount of organizational resources. Therefore, organizations could benefit greatly from accurate methods of identifying behaviorally suitable and non-suitable dogs at early ages.

Typically, professional personnel provide assessments of odor-detector dog candidates using numeric ratings that indicate how strongly the dogs express certain traits thought to be important for successful detection work. However, candidate dogs with some working organizations spend the majority of their early lives living in the homes of volunteers known as *puppy raisers.* These volunteers have extensive opportunities to observe the puppies over time and across situations. Therefore, we hypothesized that

these puppy raisers may be able to provide useful assessments of their candidate dogs' behavioral characteristics. If true, implementing a measure for obtaining accurate behavioral assessments from puppy raisers could improve an organization's ability to identify dogs suitable and not suitable for their working roles and do so at early life stages. In Study 1 (Chapter 2), we develop a trait rating survey to be completed about the puppies by puppy raisers; we then examine whether these ratings predict subsequent success in a large working-dog organization (Australian Customs & Border Protection Services; AC&BPS).

However, detector dogs do not work alone but in partnership with human *detector-dog handlers*. Together, the *dog-handler team* searches specified areas at a rapid pace. Much past research has focused on identifying personality traits and characteristics of dogs that predispose them to become effective working dogs (Maejima et al., 2007; Sinn, Gosling, & Hilliard, 2010; Slabbert & Odendaal, 1999) but the equivalent question for the human half of the dyad—whether some personality traits and characteristics might predispose a person to becoming an effective odor-detector-dog handler—has largely been ignored. A detector-dog handler is typically tasked with several responsibilities that require a specific skill set. During a typical search, the human handler follows behind a leashed dog as he/she hunts for an odor within a designated area. The handler carefully watches for behavioral cues that indicate the dog has picked up an odor. These cues can be subtle and fleeting. A handler's failure to identify a cue can easily lead to the handler-dog team's failure to detect hidden contraband. The handler must also decide whether the

dog has thoroughly searched each area and guide the dog to re-search spots he/she feels have not been satisfactorily searched.

Thus, job performance of the handler is absolutely central to the success of the handler-dog dyad, and many working-dog organizations commit a significant portion of their resources to selecting and training new dog handlers. In Study 2 (Chapter 3), we undertake an exploratory analysis on AC&BPS personnel to identify whether traits of humans can help predict future job performance as a dog handler. Study 2 tests four hypotheses addressing the relationships between dog handler job performance and handler demographics, personality, attitudes towards the human-dog relationship, and sports participation.

Chapter 2

Development of Subjective Ratings Survey for Detector Dog Puppy Raisers

BACKGROUND

Certain definable and observable traits are believed to be essential for dogs to perform detection work effectively. Traits can be thought of as relatively stable behavioral, emotional, and mental characteristics that differ between individuals (Goldberg, 1999). Therefore, dogs are often selected for detection work based on trait evaluations. The two most common methods for evaluating individual differences in traits among dogs are *codings* and *ratings* (Gosling & Vazire, 2002; Jones & Gosling, 2005; Vazire, Gosling, Dickey, & Schapiro, 2007). Coding methods count the frequencies or durations of observed behaviors by each dog during an observational period (e.g., Haverbeke, De Smet, Depiereux, Giffroy, & Diederich, 2009). For example, to behaviorally code a dog's tendency to use scent as an investigation tool, one might calculate the proportion of time spent sniffing the ground vs. not sniffing the ground during a discrete time period.

Ratings methods generally require observers to assign ratings on their overall intuitive impressions of the subjects (e.g., Gosling, Kwan, & John, 2003). There are two types of ratings methods, which differ in terms of the sample of behaviors on which the ratings are based (Svartberg, 2007). The first, behavior ratings, are the most commonly used in working-dog assessments (Jones & Gosling, 2005). Behavior ratings are based on behaviors observed during a discrete observational or testing scenario. The observer

generally attempts to be as objective as possible when assigning ratings. For example, military working dogs have been given a series of standardized tests designed to measure traits such as "confidence," "hunt drive," and "independence" (e.g. Sinn, Hixon, & Gosling, 2011), all measured on a 1 – 5 Likert scale (increasing scores mean 'more' of that trait).

One issue for both codings and behavioral ratings is that observations made only during standardized tests are inherently limited to capturing only those behaviors exhibited during the finite time periods of the tests. In other words, using codings or behavioral ratings may never capture dogs' reactions to certain stimuli, if the stimuli are not specifically presented during testing procedures.

The second type of ratings method is subjective ratings. Subjective ratings most often include a rater's overall experience with a subject rather than through specific test settings (Block, 1961). For example, a subjective rating survey might ask a rater to indicate how a dog tends to react in a variety of scenarios, such as when strangers approach or when being bathed. Subjective ratings are usually most accurate when raters know the subjects well (e.g., Gosling, 1998), and research has shown dog owners to be relatively accurate when using subjective ratings to make behavioral assessments of their pet dogs (e.g., Gosling, Kwan, & John, 2003; Hsu & Serpell, 2003; Ley, McGreevy, & Bennett, 2009). This finding is unsurprising, because one would assume that dog owners are the people who are best positioned to know their dogs, and who have the most opportunities to witness their dogs' reactions across a wide assortment of scenarios.

Behavior codings and behavior ratings are often thought to be more reliable methods of observing behavior than subjective ratings because of the latter's reliance on rater intuition. It is also true that in many cases, working dogs belong to their respective organizations. Therefore, it could be argued that working dogs do not have the benefit of having true owners that can reliably give subjective ratings. Perhaps for these reasons, subjective ratings have not been extensively used for working-dog assessments. However, empirical studies have shown that subjective ratings methods correlate substantially with both behavior ratings and behavior codings (De Meester, Pluijmakers, Vermeire, & Laevens, 2011; Gosling, Kwan, & John, 2003; Hewson, Luescher, & Ball, 1998; Van den Berg, Schilder, De Vries, Leegwater, & van Oost, 2006).

Some detector-dog programs place young candidate puppies with civilian families who raise the puppies in their homes until working outcomes have been determined. It is possible that these volunteer puppy raisers, acting as *de facto* owners, may be able to accurately rate their puppies equally as well as companion-dog owners are. Puppy raisers may lack the expertise of professional organizational personnel but there is some evidence that novices can be quite good at interpreting dog behavior and that dog owners are able to give realistic assessments of their dogs (e.g., Gosling, Kwan, & John, 2003; Hsu & Serpell, 2003). Therefore, it is plausible that puppy raisers could give information based on subjective ratings that would help determine behavioral suitability and non-suitability for work. Several researchers have examined the ability of subjective ratings used in test batteries to predict success in working dogs (see Sinn, Gosling, & Hilliard,

2010 for review), however, little is known concerning whether subjective ratings made by puppy raisers can accurately predict working-dog outcomes.

The goal of this study was to develop and administer an effective puppy raiser subjective rating survey for households in Melbourne, Australia, currently volunteering as puppy raisers for the Australian Customs and Border Protection Service (AC&BPS). The AC&BPS breeds Labrador Retriever puppies that are trained to detect illegal drugs and explosives in environments such as airports and mail centers. AC&BPS puppy raisers generally receive their puppies at the age of 8 weeks and raise them until they are up to 15-months old. AC&BPS personnel perform near-monthly behavior tests on puppies at their Melbourne based facility, or near the puppy raisers' homes. Dogs may be eliminated from candidacy at any stage if deemed not suitable. At ages ranging from 12 months to 15 months, non-eliminated dogs are given final assessments. Successful dogs display the traits suitable to become working dogs, and graduate to work with the AC&BPS or are sold to work for other organizations. Non-successful dogs, who do not display suitable working-dog traits, are sold to civilian homes as pets.

To be useful, the Puppy Raiser Subjective (PRS) Survey would need to demonstrate reliability across independent raters (inter-rater reliability; Sinn et al., 2010) and be able to accurately predict the work-related outcomes of the dogs (predictive validity; Sinn et al., 2010). We conducted the current study in two phases. In Phase 1, we selected trait items and developed the PRS Survey. In Phase 2, we implemented the survey on AC&BPS puppy raisers and analyzed the results for inter-rater reliability and

for underlying factor structures of the survey items. We then tested the results for predictive validity.

PHASE I: SURVEY DEVELOPMENT

In selecting items for the PRS Survey we sought to achieve two goals. The first goal was to ensure that as many potentially relevant behavioral dimensions as possible were included in the instrument. The second goal was to minimize the amount of time required of the puppy raisers. To meet these two goals, we undertook a four-step item-selection process.

In Step 1, we developed a draft list (DL) of 26 dog behavioral traits based on our own ideas about what could be important for detector-dog work. This list was generated based on discussions with professional detector-dog handlers and our own experiences.

In Step 2, we derived a primary framework for the PRS survey using the Military Working Dog Trainer -Trait Rating Survey (MWD Survey; Gosling, 2009). The MWD Survey was designed for military-working-dog trainers as a method of assessing suitability of dogs for military work. The MWD Survey was a logical starting framework for our survey because odor detection is one of the tasks required of military-working dogs. The MWD Survey consists of 14 individual trait items and two overall work performance items. We adapted 10 items from the MWD Survey that tapped traits from our DL. Two of the MWD Survey items, "sociability towards people" and "sociability towards dogs," were later partitioned into additional items to make distinctions between levels of sociability towards unknown and familiar people and dogs.

Past research suggests that guide-dog puppy raisers may be able to accurately predict their dogs' overall chances for success (Batt, Batt, Baguley, & McGreevy, 2009); therefore, in Step 2 we also included an overall work performance item, "Do you think this dog will make a good detector dog?". Sixteen traits from the DL remained after these first two steps in the development of the PRS survey.

In Step 3 we located items from existing dog-trait surveys that could sufficiently measure the 16 remaining traits. Our aim was to use measures that had been previously validated in past research to maximize reliability and validity of PRS Survey.

The Canine Behavioral Assessment and Research Questionnaire (C-BARQ; Hsu & Serpell, 2003) was the first instrument we examined in Step 3. The C-BARQ is a previously validated comprehensive questionnaire containing 101 items measuring 13 individual traits in companion dogs (e.g. trainability, separation-related behaviors). The C-BARQ instrument has been shown to have strong reliability and validity in studies spanning across multiple types of dog populations. Through factor analysis, Liinamo, van den Berg, Leegwater, van Arendonk and van Oost (2006) developed a shortened version of the C-BARQ (sC-BARQ; Liinamo et al, 2006) that consolidated the original 101 items into 11 items that each measured a distinct trait. For example, 27 aggression items on the long form C-BARQ capture a wide range of scenarios where aggression is commonly exhibited. The shortened sC-BARQ condenses these into four items distinguishing between aggression towards familiar and unfamiliar people and dogs. Eight items (four "aggression" items, "prey drive," "separation anxiety," "excitability," and "trainability") that were relevant to the detector dog population were adapted from the C-BARQ and sC-

BARQ for the PRS Survey.

The Dog Personality Questionnaire (DPQ; Jones, 2008) was the second instrument we examined in Step 3. The DPQ is a 75-item survey for companion dog owners measuring five individual traits with each trait having multiple facets. The DPQ has previously been rigorously tested for reliability and validity in a series of six studies (Jones, 2008). Items making up the DPQ trait facet "playfulness" were adapted into a single item for use in the PRS Survey. Items that addressed the tendency for a dog to chase moving objects like bicycles and skateboarders were adapted from the DPQ trait facet "prey drive" to supplement the sC-BARQ derived "prey drive" item, which had limited itself to the tendency to chase only live prey animals like squirrels and cats. Two PRS Survey items ("nosy/uses nose" and "stamina") were derived from an already existing third instrument surveying Air Force, Army, Marine Corps, and Navy military working dog handlers on their opinions of behavioral traits that were valuable to explosive detection dogs (Jones, Sinn, Gosling, & Hilliard, 2011, 2012). We combined the term "intelligence" identified by the handlers with the item measuring "trainability" that was derived from the C-BARQ and sC-BARQ.

Two traits from the original 26-item DL ("food motivation" and "kennel behavior") could not be found on previous questionnaires. We speculated that strong "food motivation" could cause differences in how detector dogs respond to training. For example, dogs in the current study dogs are originally taught to search for items soaked in a popular pungent food paste called Vegemite®. Therefore, its possible that highly food motivated dogs would be more focused when searching. So we created a new item to tap

this trait.

When dogs are brought into the AC&BPS for behavior testing, they are kept in individual kennels. We speculated that differences in kennel behavior might result in differences in test battery performance. For example, it is possible that frantic behavior in a kennel immediately before a test battery may affect performance differently than calm behavior before a test. So we created a new item to tap this trait.

The item-selection process resulted in a final PRS Survey consisting of 27 items (26 putative discrete traits and 1 broad item). The survey concluded with a free-text response box asking the rater if they felt any additional behavioral traits should have been included in the survey. Wording for each PRS Survey item attempted to clearly and comprehensively describe each trait using a combination of lay-person and dog-owner language. All items were formatted to remain consistent in language with the items derived from the MWD Survey.

The final version of the PRS Survey is shown in Appendix B. Each trait item was accompanied by a brief explanation and examples of what might constitute low and high scores for that trait. Traits were scored using a 7-point Likert scale (1 = *Trait is extremely uncharacteristic of this dog*, 7 = *Trait is extremely characteristic of this dog*). A not applicable ("N/A") option was included because it was felt that some puppy raisers might not have had the opportunity to witness a certain behavior or the item might not apply to the AC&BPS detection-dog population. Basic demographics about the raters and dogs were also collected, and the raters were also asked how well they knew the dogs (rated on a 1-7 Likert scale, with 7 representing *knowing the dog very well*). Before distributing the

PRS Survey to puppy raisers, three active detector-dog professionals contributed final minor modifications.

PHASE 2: INTER-RATER RELIABILITY, FACTOR STRUCTURE, AND PREDICTIVE VALIDITY OF THE PRS SURVEY

Subjects

Our dog subjects were 105 Labrador Retriever dogs (60 male, 45 female) from the AC&BPS detector dog program. The majority of the dogs were actively in the process of development and testing for detector work. Other dogs were either actively working in the field or retired dogs that had been used for breeding purposes. Dogs ranged in age from 3-months old to 8-years old ($M = 16.28$, $SD = 17.54$, $Mdn = 10$).

Raters

Human raters were members of 81 Australian households in the vicinity of Melbourne, Victoria, Australia currently raising puppies for AC&BPS detection work during the period of May, 2012 to June, 2012. A total of 208 households were given surveys throughout this period, so the overall return rate at the level of households was 39%. Eighty-one households returned at least two surveys for an AC&BPS dog in their home. A total of 186 independent human participants returned surveys for at least one dog. Twenty households returned surveys for two puppies that they were raising and three households returned surveys for three puppies. Human raters ranged from 9 to 77 years old ($M = 40.7$, $Mdn = 44$, $SD = 15.9$). The total sample consisted of 105 females (*M*

age = 37.69, *Mdn* = 42, *SD* = 16.33) and 81 males (*M* age = 42.11, *Mdn* = 45, *SD* = 15.99).

Procedure

Households were given paper copies of the PRS Survey, a written description of the study (Appendix C), and an instructional DVD were combined in a survey packet for each household. Households caring for more than one dog were given additional surveys. All packet materials were enclosed in a pre-addressed, postage paid envelope for return to the AC&BPS facility. Survey packets were primarily delivered directly to the households by AC&BPS officers during routine visits and scheduled pick-ups. A small number of surveys were given directly to household members visiting the AC&BPS facility. Survey packets were primarily returned via mail; some households returned surveys directly to customs officials returning dogs from scheduled evaluations.

Four paper copies of the PRS survey were given to each household; instructions asked that surveys be completed for each dog by up to four separate individuals. Only individuals in the household that were familiar with the dog(s) were asked to provide ratings. Raters were informed that ratings should be based on their overall familiarity with the dogs, using their personal understanding of the traits, included trait explanations, and understanding of typical detector-dog behavior as guides. They were instructed to select the "N/A" option if they had never been given the opportunity to witness a trait or did not understand a trait. Participants were asked to make their ratings independently of any other household members, and not discuss the traits of the dogs prior to or during the completion of the survey.

Descriptive Statistics

Descriptive statistics for individual survey items are shown in Table 2.1. For traits where high scores indicated clearly desirable characteristics (e.g., "intelligence/trainability," "willingness to please," all four "sociability" items) raters mainly used the upper half of the scoring scale. For traits where high scores indicated clearly undesirable characteristics (e.g., "aggression" items, "separation anxiety," "kennel behavior"), raters primarily used the lower half of the scale. These distributions indicate that overall, raters felt that their dogs possessed characteristics making them suitable for detection work. Descriptive characteristics of the broad evaluation item explicitly asking raters whether the dog would make a good detector dog further support this sentiment ($M = 5.95$, $SD = 1.17$).

Only two of the questionnaire items ("independent," "kennel behavior") had proportions of N/A scores that neared 20% (Table 2.1), indicating that human participants had difficulty relating these items to their dogs.

The survey concluded with a free text box for rater comments, which is given in Appendix D.

Table 2.1. Descriptive statistics for the item responses from the PRS Survey.

Item	N	Mean	SD	Range	Skew*	Kurt*	% of N/A scores
Intelligence/Trainability	105	5.44	1.16	5	-0.60	-0.23	0.00
Willingness to Please	105	6.18	1.04	5	-1.49	2.30	0.00
Environmental Soundness	105	5.86	1.14	5	-0.98	0.61	0.00
Desire for toy, towel, or ball reward	105	6.00	1.26	6	-1.53	2.37	0.00
Search/Hunt Drive	99	5.30	1.42	6	-0.71	0.03	8.00
Focus/Determined	104	5.13	1.18	6	-0.61	0.48	4.20
Independent	94	4.84	1.40	6	-0.56	-0.04	**21.43**
General Liveliness/Energy Level	105	6.16	1.07	5	-1.43	1.98	0.00
Hardness	105	5.65	1.26	6	-1.18	1.41	2.80
Curiosity	105	6.14	1.00	5	-1.25	1.46	1.00
Excitability	105	4.74	1.38	6	-0.47	-0.03	0.00
Food Motivation	104	5.70	1.23	5	-0.71	-0.22	1.40
Nosiness	105	6.22	1.07	5	-1.44	1.60	0.00
Stamina	105	6.06	1.13	5	-1.33	1.56	1.00
Playfulness	105	6.38	0.93	5	-1.93	4.56	1.00
Kennel Behavior	98	3.04	1.68	6	0.36	-0.97	**18.10**
Sociable - unknown people	105	5.65	1.10	5	-0.87	0.77	0.00
Sociable - unknown dogs	105	5.96	1.05	5	-0.85	0.14	0.00
Sociable - unknown dogs	105	6.02	0.93	4	-0.64	-0.35	1.00
Sociable - unknown people	102	6.16	0.90	5	-1.14	1.76	1.00
Aggression - familiar people	105	1.84	1.73	6	2.20	3.49	2.30
Aggression - familiar dogs	101	1.84	1.63	6	2.28	4.13	9.00
Aggression - unknown people	105	1.74	1.67	6	2.45	4.65	2.80
Aggression - unknown dogs	104	1.85	1.57	6	2.24	4.12	4.70
Prey Drive	105	4.18	1.95	6	-0.16	-1.16	1.40
Separation Anxiety	105	2.74	1.59	6	0.65	-0.63	1.00
Do you think this dog will make a good detector dog?	103	5.95	1.17	6	-1.43	2.29	2.80

[1] There are two types of kurtosis (kurt), leptokurtotsis (L) and platykurtosis (P). The former is when there a distribution has an acute peak around the mean (positive kurtosis values), while the latter indicates a lower, wider peak around the mean (negative kurtosis), both relative to a normal distribution. Similarly, a negative skewness (skew) indicates that the tail to the left of the probability density function is longer than the right, and a positive value indicates the reverse. In practice, values of skewness and kurtosis that are close to zero indicate a normal distribution. Extremely skewed variables can usually be transformed to fit closer to a normal distribution. Leptokurtosis (i.e., a particular type of lack of variation) is problematic for statistical analyses, since it cannot be dealt with via variable transformation. 'Playfulness' and the four items related to 'aggression' were all extremely skewed and kurtose (most raters responded in the same way per item, all giving relatively the same large or small score (extreme values of skewness and kurtosis in bold).

Inter-rater reliability

We examined the level of agreement between raters. Inter-rater reliability was evaluated using the Intra-Class Coefficient (ICC). The ICC uses analysis of variance (ANOVA) to compare the proportion of the overall data variance caused by raters to the proportion of the variance caused by the subjects (Gwet, 2001; Shrout & Fleiss, 1979). If subjects are the main cause of variance, with raters contributing very little, then an item will show a strong ICC estimate, indicating agreement between raters is strong. If the raters are the main cause of variance, with subjects contributing very little, then an item will show a weak ICC estimate, indicating less agreement between raters in their assigned ratings. We used generally accepted practices to determine whether inter-rater reliability was strong, moderate, or weak. An ICC estimate of 0.70 or above was considered an indicator that an item had strong inter-rater reliability. An ICC estimate of 0.50 to 0.69 was considered an indicator of moderate inter-rater reliability, and an ICC estimate below .050 was considered an indicator of weak inter-rater reliability.

Multiple types of ICC's exist, each type appropriate for differing data characteristics. Our study consisted of separate pairs of raters, with each pair scoring individual dogs. Therefore, a one-way ANOVA was used in our study, as suggested by Shrout and Fleiss (1979). Average ICC estimates were interpreted which provides reliability estimates for each item based on the average of all raters.

Table 2.2 shows each trait partitioned into categories of high, moderate, and low inter-rater reliability. Using data from all dogs, inter-rater reliability estimates for all items ranged from 0.33 (L95% C.I. = 0.10, U95% C.I. = 0.55) for "curiosity" to 0.81 (L95% C.I. = 0.72, U95% C.I. = 0.87) for "Do you think this dog will make a good detector dog?"; the mean reliability estimate of all items was 0.63 (L95% C.I. = 0.43; U95% C.I. = 0.72). "Do you think this dog will make a good detector dog?," "nosy/uses

nose," "prey drive," "environmental soundness," "search/hunt drive," "willingness to please," "stamina," "independent," "desire for toy," and "intelligence/trainability" all had ICC estimates >= 0.70, indicative of strong inter-rater reliability. ICC estimates of three further items, "focus/determined/task oriented," "general liveliness/energy level/activity level," and "kennel behavior," were slightly below the 0.70 threshold (0.69, 0.68, 0.68, respectively), indicating borderline strong inter-rater reliability.

A further nine items were considered to have moderate inter-rater reliability (ICC estimates 0.53 to 0.65). "Aggression towards familiar people," "aggression towards unknown dogs," "aggression towards familiar dogs," "playfulness," "sociable towards unknown dogs," "separation anxiety," "sociable towards familiar dogs," "food motivation," and "excitability" all had estimates that were indicative of moderate agreement amongst observers. "Sociable towards unknown people" indicated borderline moderate inter-rater reliability (ICC = 0.49).

"Curiosity," "sociable towards familiar people," "aggression towards unknown people," and "hardness" showed the weakest reliability, all scoring in ranges considered weak (ICC estimates <= 0.43).

Table 2.2. Intra-class correlation coefficients (ICCs), 95% confidence intervals, and inter-rater reliability (IRR) inferences for PSR survey items

SR Item	ICC's	95% CI	N	Inference
Do you think this dog will make a good detector dog?	0.81	.72 - .87	101	Strong IRR
Nosy/uses nose	0.76	.64 - .84	105	
Prey Drive	0.73	.59 - .81	102	
Environmental Soundness	0.72	.59 - .81	105	
Search/Hunt Drive	0.72	.58 - .81	95	
Willingness to Please	0.72	.59 - .81	105	
Stamina	0.72	.58 - .81	104	
Independent	0.71	.53 - .82	72	
Desire for toy, towel, or ball reward	0.70	.55 - .79	105	
Intelligence/Trainability	0.70	.55 - .79	105	
Focus/Determined/Task Oriented	0.69	.54 - .80	97	Moderate IRR
General Liveliness/Energy Level/Activity Level	0.68	.30 - .68	105	
Kennel Behavior	0.68	.50 - .80	74	
Human aggression – towards familiar people	0.64	.47 - .76	100	
Dog aggression - towards unknown dogs	0.63	.44 - .75	96	
Dog aggression - towards familiar dogs	0.61	.40 - .74	91	
Playfulness	0.58	.38 - .72	103	
Sociable - towards unknown dogs	0.57	.36 - .71	104	
Separation Anxiety	0.56	.35 - .70	103	
Sociable - towards familiar dogs	0.54	.30 - .70	88	
Food Motivation	0.54	.32 - .69	103	
Excitability	0.53	.30 - .68	105	
Sociable - towards unknown people	0.49	.25 - .65	105	Weak IRR
Hardness	0.43	.16 - .62	99	
Human aggression – towards unknown people	0.43	.15 - .61	99	
Sociable - towards familiar people	0.35	.04 - .57	104	
Curiosity	0.33	.10 - .55	103	
Average*	0.63	.43 - .72		

* average estimated by transforming all ICC estimates to Fischer's z values, averaging them, then backtransforming the resulting average z to an average r

Principal Components Analysis

We used principal components analysis (PCA) using the correlation matrix of the individual items to determine whether our survey measured major underlying dimensions (Jolliffe, 2002). The sampling adequacy of the correlation matrix was assessed using the Bartlett sphericity test (χ^2= 314.08, df = 325, p = 0.66) and the Kaiser-Meyer Olkin measure (KMO = 0.73; Budaev, 2010). So that variance of all 26 single trait items could contribute to the PCA, no items were eliminated due to weak inter-rater reliability. The overall performance item "Do you think this dog will make a good detector dog?" was not included in the PCA because it was not intended to measure any single distinct trait.

An oblique rotation (Promax) was used to permit expected correlations to exist between factors. By combining variables into broader factors, we undoubtedly lose information about the finer points the PRS Survey items were attempting to measure. However, it is much more efficient when deriving predictive models if a smaller number of variables can be used without the loss of significant information.

PCA results were examined in multiple ways to determine the best factor structure, including the Kaiser rule (Kaiser, 1960), Cattell's scree test (Cattell, 1966), and Goldberg's "bass-ackwards" hierarchical method (Goldberg, 2006). We also used the interpretability of the factors in each of the various solutions to reach a final solution.

The Kaiser rule states that any factor with an eigenvalue over 1.0 should be retained, suggesting that we should retain seven factors. Cattell's scree test plots the eigenvalue of each factor, which can be thought of as an index of the amount of variance explained by the factor, along a y-axis. An imaginary line drawn from left to right through each eigenvalue will often reveal a noticeable break where the line shifts from sloping downwards to being virtually flat. Cattell's test suggests that only the factors to the left of this "elbow" should be retained. The resulting scree plot from our data shows a

clear elbow suggesting a three-factor solution (Figure 2.1).

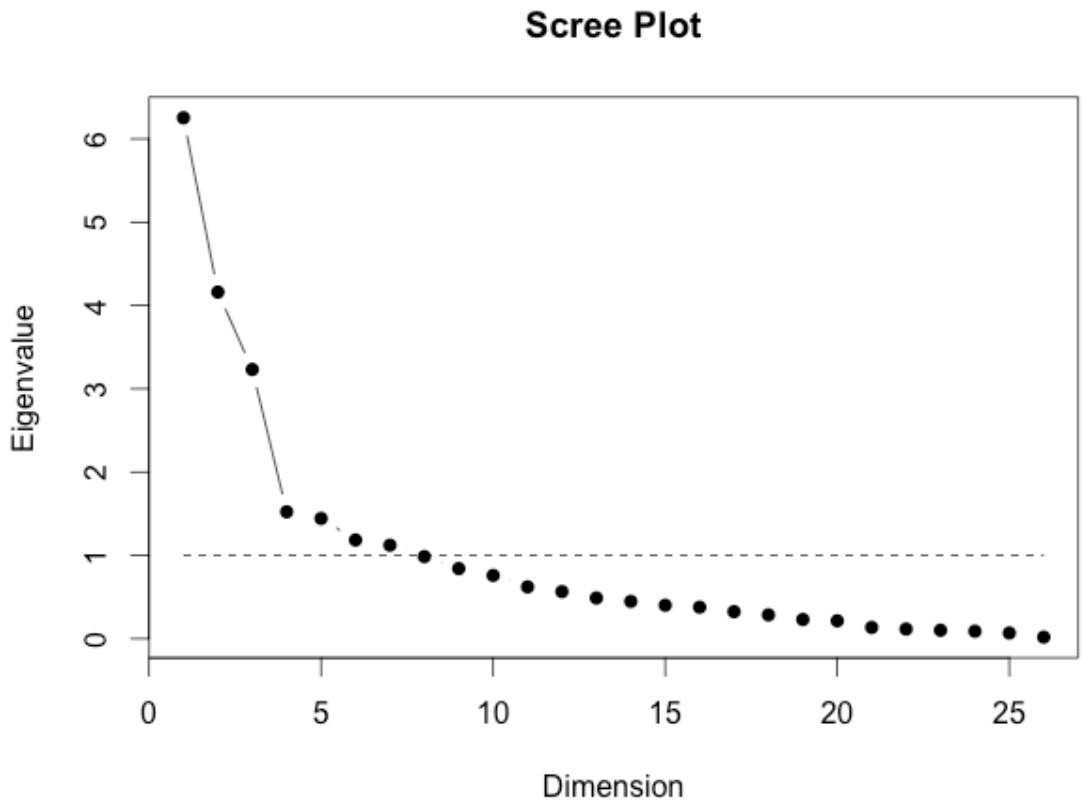


Figure 2.1. Scree plot of covariance matrix. Points on the plot indicate the eigenvalue (y-axis) of each factor dimension (x-axis). The dotted horizontal line shows the eigenvalue of 1. According to the Kaiser rule, any point above this line should be retained as a factor. Cattell's scree test shows a distinct elbow at the fourth factor dimension, suggesting a 3-factor solution. Less distinct elbows can be identified subsequent points, which could contribute to the discrepancies of our solutions

We also implemented Goldberg's *bass-ackwards* hierarchical method for the number of factors to extract/interpret in a PCA (Goldberg, 2006). The bass-ackwards method is a "top-down" approach that starts with a single, general factor. Factors of increasing specificity are then extracted until a clear, interpretable factor structure is determined (Goldberg, 2006). Examining the factors at each hierarchical level reveals the consequences of extracting too many or too few factors. At each hierarchical level, labels are applied to the factors that attempt to capture the primary commonalities of their respective items. Figure 2.2 shows a diagram of our bass-ackwards analysis. The first three factors extracted included items that strongly suggested broader dimensions we called *Trainability*, *Aggression*, and *Sociability* respectively. These three factors remained largely unchanged with subsequent extractions. Items in the fourth factor were less interpretable, but suggested characteristics common to *Confidence*. Subsequent factors primarily pulled items from Confidence, resulting in an additional three interpretable factors: *Energy Level*, *Food Drive,* and *Emotional Arousal*. An eighth factor began breaking apart factors illogically, creating a less interpretable solution. We concluded that a seven-factor solution (Table 2.3), consisting of four major factors (Trainability, Aggression, Sociability, Confidence) explaining 51% of the variance and three minor factors (Energy Level, Food Drive, Emotional Arousal) explaining an additional 30%, gave the clearest interpretability. This solution also seemed to explain the discrepancies found between solutions suggested by the Kaiser test and Cattell's test.

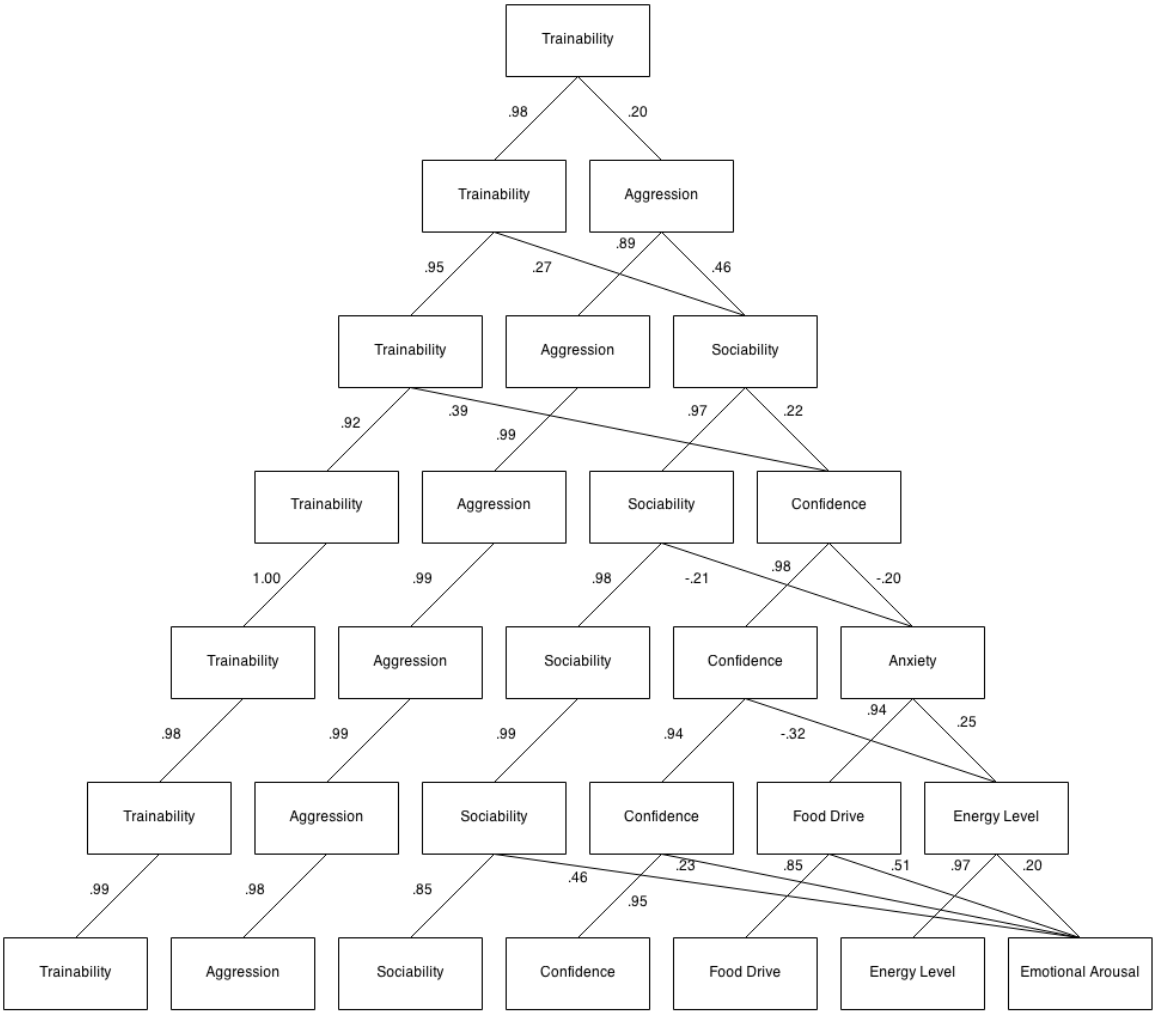


Figure 2.2. Bass-ackwards hierarchical factor analysis. The first factor extracted at the top was the Trainability factor. At each descending level, an additional factor is extracted. The bottom level shows the final 7-factor solution. For clarity, only correlations above 0.20 are shown.

Table 2.3. Factor Loadings of 26 Dog Trait Items on Seven Promax-Rotated Principal Components

Trait items	Trainability	Aggressiveness	Sociability	Energy Level	Food Drive	Confidence	Emotional Arousal
hunt drive	**0.99**	0.11	0.04	-0.21	0.00	-0.06	-0.06
focus	**0.95**	-0.12	-0.05	-0.08	0.01	0.06	0.13
toy drive	**0.84**	-0.04	-0.09	-0.02	0.17	-0.04	-0.02
intelligence	**0.79**	0.03	0.07	0.09	-0.43	0.04	0.06
will. To please	**0.72**	0.04	0.11	0.00	-0.23	0.05	0.00
playfulness	**0.67**	0.02	-0.01	0.18	0.22	-0.17	-0.16
independence	**0.56**	0.07	0.04	0.09	0.06	0.15	-0.14
agg – unk. people	0.04	**0.97**	-0.03	0.00	0.03	-0.09	-0.02
agg – unk. dogs	0.08	**0.96**	0.01	-0.10	0.01	-0.10	-0.03
agg – fam people	-0.02	**0.95**	0.01	0.06	0.01	0.09	0.00
agg – fam. dogs	-0.08	**0.94**	-0.05	0.02	0.15	-0.07	0.05
soc – fam. dogs	0.01	0.03	**0.96**	-0.24	0.00	0.24	-0.08
soc – unk. dogs	0.13	-0.14	**0.80**	-0.22	0.27	0.05	-0.04
soc – fam. people	-0.01	0.03	**0.69**	0.13	-0.03	-0.26	0.13
soc – unk. dogs	-0.24	0.01	**0.59**	0.37	0.05	-0.10	-0.04
general liveliness	-0.10	-0.05	-0.06	**0.91**	0.02	0.10	-0.03
stamina	-0.04	0.05	-0.13	**0.84**	-0.17	0.27	0.02
food motivation	-0.03	0.07	0.09	-0.24	**0.78**	0.03	0.15
nosy/uses nose	-0.05	0.12	0.09	0.29	**0.65**	0.16	0.09
hardness	-0.16	-0.09	0.09	0.30	0.01	**0.87**	-0.09
env. soundness	0.35	-0.06	-0.09	0.01	0.17	**0.69**	0.18
curiosity	0.16	0.07	0.09	**0.39**	0.31	**0.37**	-0.05
separation anxiety	-0.25	0.03	-0.02	-0.24	0.25	0.12	**0.69**
prey drive	0.20	-0.11	-0.14	0.17	0.26	-0.20	**0.66**
kennel behavior	-0.04	0.13	0.07	-0.09	-0.38	0.23	**0.59**
excitability	0.14	-0.03	0.23	0.33	-0.06	-0.22	**0.45**
% of variance explained	0.19	0.15	0.10	0.09	0.07	0.07	0.06
Cronbach's alpha	0.90	0.97	0.80	0.74	0.71	0.56	0.46

Notes. Loadings in bold font indicate strongest item loadings. Crombach's alpha measures internal consistency of items included in a factor. Factors with alphas closer to 1.0 indicate that the items included are more reliably measuring the same behaviors, and factors closer to 0.0 indicate that the items included are less reliably measuring the same behavior. Diagnostics for PCA were conducted using Bartlett's test (

Each dog received an estimated factor score corresponding with each of the seven extracted factors. Factor scores were calculated by averaging the single trait item scores loading most strongly onto a factor. As would be expected, several items cross-loaded onto more than one factor. The survey item "curiosity" loaded nearly equally on Trainability, Energy Level, Confidence, and Food Drive. It is likely that a dog's curiosity could be demonstrated in multiple ways that could easily be related to any of these factors. For example, a dog that is highly energetic may often be investigating the surroundings, searching for things to do or eat. This could be interpreted as "curiosity" by the rater, explaining the positive loadings on both Energy Level and Food Drive factors. We chose to include "curiosity" in the Confidence factor for two reasons: 1) in AC&BPS behavior test batteries, the trait "curiosity" contributes to a score on a broad category similarly labeled Confidence, and 2) when extracting more than seven factors, curiosity continued loading onto the Confidence factor, while its loadings on other factors decreased.

The item "nosy/uses nose" was intended to be used as a measure of a dog's tendency to rely on his/her nose when searching for an object, and therefore might seem peculiar in a Food Drive factor. We speculate two possible reasons for this loading. One possibility is that dogs are initially trained using Vegemite® paste, so the smell of food and their ability to use their noses are directly related. The second possible reason is that while professional handlers are actively trying to identify dogs that rely on their noses instead of other senses when performing searches, the average puppy raiser is probably not actively looking for this trait and it is therefore not highly noticeable. Therefore, it is likely that raters might rate "nosy" based on how strongly dogs respond when food is present. Interestingly, "nosy" was the trait with the highest inter-rater reliability, and no

raters selected N/A for the item. This indicates that raters felt they understood the item and tended to agree with one another.

Four items, "separation anxiety," "kennel behavior," "prey drive," and "excitability," all seemed to contrast levels of calm behavior vs. hyperactive behavior in the presence of differing stimuli; we named this factor Emotional Arousal. High emotional arousal could be characterized by exhibiting hyperactive behavior when faced with certain stimuli, and low emotional arousal could be characterized by exhibiting calm behavior in the presence of those same stimuli. It is likely that a dog that acts hyperactively when left alone, suggesting high scores for separation anxiety, will also act hyperactively when left alone in a kennel, suggesting high scores for kennel behavior. High scores in these two items would also sensibly result in high scores in the Emotional Arousal factor. In addition to loading on the Emotional Arousal factor, the excitability item also loaded onto Sociability and Energy Level factors. These loadings make sense as raters might view dogs that are generally hyperactive or display hyperactivity in the presence of people or dogs as excitable, and therefore, high in Emotional Arousal. Additionally, dogs may act hyperactively when they see squirrels or cats, leading to high scores in "prey drive". Therefore, "prey drive" is also easy to associate with high Emotional Arousal.

We examined internal consistency of the factors using Cronbach's alpha (Table 2.2), which indexes how strongly items loading onto a single factor are all measuring the same construct. Trainability (α= 0.90) and Aggression (α= 0.97) showed strong internal consistency, while the consistencies of Sociability (α= 0.80), Confidence (α= 0.74), and Energy Level (α= 0.71) were only moderate. Food Drive (α= 0.56) and Emotional Arousal (α= 0.46) showed low internal consistency. We next examined whether internal consistency of any factors could be improved by removing items. Only the item

"independent" had a negative impact on its corresponding factor, with results showing that its removal would lead to a 0.01 alpha increase for Trainability. Because of this relatively mild impact, we chose to retain "independent" in the analysis.

Predictive Validity

In Phase 4 we examined predictive validity of our factors on the success and non-success of dog outcomes. One PRS Survey item, "Do you think this dog will make a good detector dog?," was intended to provide an overall assessment of each dog by the rater. To analyze this item, we entered it into a logistic regression as a predictor of dog outcome, separately from other survey items, but found no relationship between puppy raisers' overall assessment scores and the true working outcomes for the dogs ($b = 0.24, p = 0.18$). Outcomes for 31 dogs over 15 months of age in our sample were likely to have already been determined at the time the survey was administered. We ran a linear regression model to examine whether the pre-determined outcomes may have affected overall assessment scores and results were not significant ($b = 0.02, p = 0.69$).

We examined internal predictive validity using cross validation on the collected sample results. In cross validation, a random portion of the data is set aside for use as a testing set. The remainder of the data is used as a training set. A predictive model is developed based only on the data contained in the training set. That model is then used to predict outcomes in the testing set. There are multiple methods of splitting data into training and test sets. We selected the Leave One Out Cross Validation (LOOCV) method so that we could use the maximum number of cases in the training set (Hawkins, Basak, & Mills, 2003). The LOOCV uses all except one of the cases in a sample to make a prediction on the excluded case. This process is then repeated until every case has been given a prediction based on the information from the other cases. For example, using our

data, the model would put “Dog 1” into the test set using the information from other dogs (training set) in the sample. Then, it would make a prediction on “Dog 2” based on the other dogs, which would now include “Dog 1”. The model continues to run until all dogs have received predictions.

Model selection and LOOCV analysis was performed using the Train function in the R add-on ‘caret’ package (Wing et al., 2014). We first conducted an LOOCV analysis on a logistic regression model featuring all seven factors, dog age, and dog gender, as predictors of successful or non-successful dog outcomes. We used the method = “glmStepAIC” automatic stepwise selection feature included in the caret package, running in both directions, to determine which combination of the seven predictors resulted in the best fitting model. The Akaike Information Statistic (*AICAIC*) statistic is used at the criteria for including each predictor. A model with a lower *AICAIC* indicates a better fitting model. The stepwise selection process evaluates each predictor, and eliminates the one that will result in the largest reduction in the *AICAIC*. After a predictor is removed, the previously eliminated predictors are rescanned to determine whether adding them back to the model results in a lower *AIC*. This process is repeated until removing any further predictors causes the *AIC* to increase.

Results of the LOOCV analysis indicated that the best fitting model featured main effects of Trainability and Sociability as predictors of dog outcome (*AICAIC* = 119.88). Dropping Sociability and leaving Trainability as the only predictor in a second model resulted in only a slightly inferior model (*AIC* = 120.55), meaning that Trainability predicted outcomes equally well or better without Sociability. This finding was confirmed using a likelihood ratio test that showed there were no significant difference between the Trainability and Sociability model and the Trainability only model (p = 0.27). In other words, Sociability did not provide significant predictive value on top of

Trainability. A third model was examined to test for an interaction effect between Trainability and Sociability but no evidence of an interaction was found (*AIC* = 121.64). Therefore, we determined that the model best predicting dog success and non-success featured only a main effect of the Trainability factor. The odds ratio of this model showed that for every unit increase in Trainability, the odds of a Pass prediction increase by a factor of 1.94, representing a 94% increase.

We ran a second LOOCV analysis using the Trainability only model to extract predictions for each dog in our sample. We were unable to obtain outcome information for one dog. Therefore, for each round of LOOCV, data from 103 dogs was used to predict the probability of the 104th dog becoming a successful detector dog or breeder (Pass outcome). If the predicted probability of a dog becoming successful was between .50 and 1.0, the dog was labeled as a Pass outcome. If the predicted probability of a dog becoming successful was between 0.00 and 0.49, the dog was labeled as a Fail outcome. Predicted probability of a Pass outcome as a function of one-unit increases of Trainability score are shown in Figure 2.3, with a dotted horizontal line representing the 0.50 threshold that must be exceeded for a dog to receive a Pass outcome label. This figure shows that all Trainability scores above approximately 4.14 were labeled as Pass outcomes. Probabilities of Pass outcomes for all dogs are shown in Appendix E.

We then compared our model's predicted Pass/Fail outcomes to the actual real life outcomes of each dog. Dogs that had become a working dog or breeder for any organization were considered to be successful. Dogs who had been sold to the public as pets were considered non-successful. Overall, our model (Table 2.4) was able to accurately predict dog outcomes 69% of the time. However, the accuracy statistic is not sufficient for making conclusions about the model because it only accounts for predictions of successful outcomes, not unsuccessful outcomes. A good predictive model

would need to correctly label successful dogs with Pass outcomes and correctly label non-successful dogs with Fail outcomes. Of the 104 dogs in our sample, our model labeled 93% as Pass outcomes (Figure 2.4). The proportion of the sample that actually became successful detector dogs was 72%, considerably less than the 93% predicted. Therefore, the 69% accuracy rate is likely to be the result of the model predicting Pass outcomes at an excessive rate. To demonstrate this phenomena, imagine a hypothetical scenario where the model predicted only Pass outcomes for 100% of the 104 dogs. If the actual percentage of successful detector dogs was 69%, then the model would have predicted accurately 69% of the time because all actually successful dogs were correctly classified as Pass outcomes, while 0% of actually non-successful dogs were correctly classified as Fail outcomes (31% of the dogs). This hypothetical model would not be useful because it was unable to distinguish between Pass outcomes and Fail outcomes.

Table 2.4. Logistic regression output showing estimated changes in predicted probability as a function of Trainability score.

	b	*SE*	*odds ratio*	*p*
Intercept	-2.73	1.53	0.07	0.08
Trainability	0.66	0.27	1.93	0.02
$\chi^2(df)$	123.11(103)			
AIC	120.62			

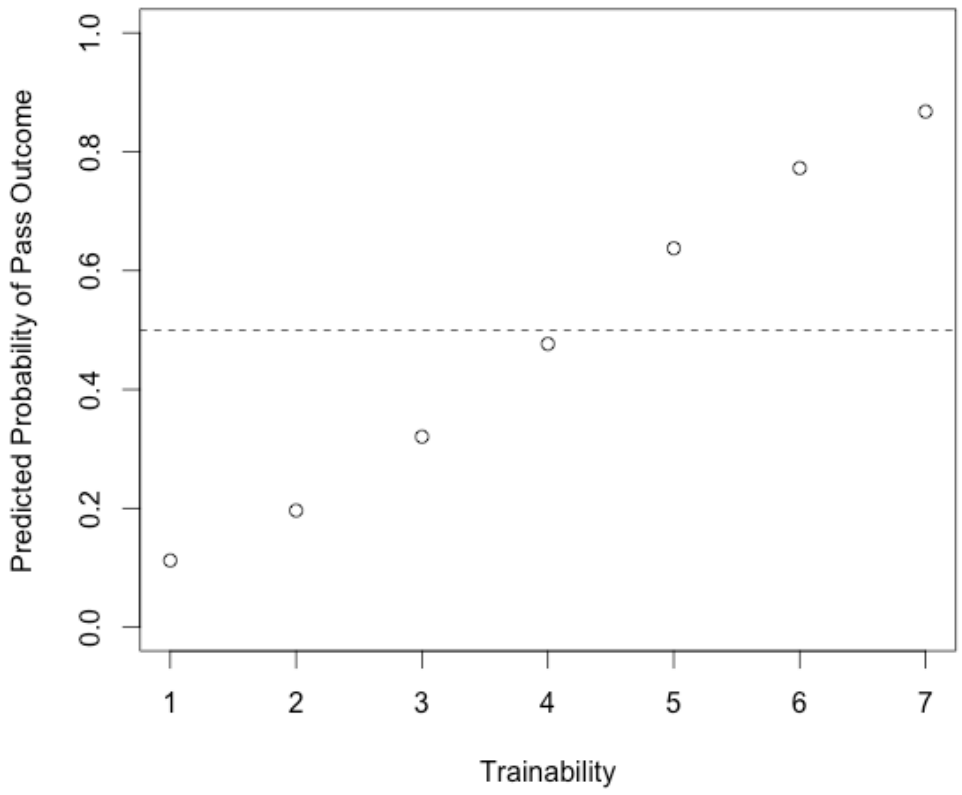


Figure 2.3. Predicted probabilities of Pass outcomes at one-unit increases of Trainability scores. Probabilities above the 0.50 threshold, indicated by the dotted line, were assigned Pass labels. Probabilities below the 0.50 threshold were assigned Fail labels.

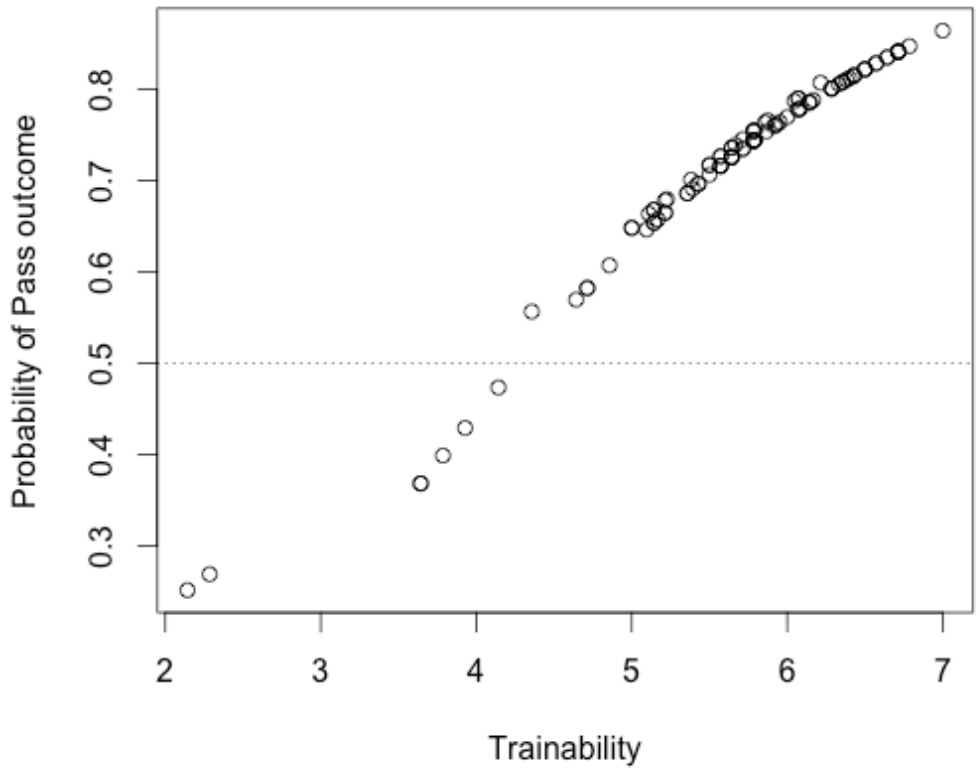


Figure 2.4. Pass probabilities predicted for each dog by Trainability factor. Probabilities above the 0.50 threshold, indicated by the dotted line, were assigned Pass labels. Probabilities below the 0.50 threshold were assigned Fail labels.

For every round of LOOCV, a predicted Pass or Fail label falls into one of four discrete categories (Table 2.5). The prediction can be a True Positive (Pass label correctly given to successful dog), a False Positive (Pass label incorrectly given to non-successful dog), a True Negative (Fail label correctly given to non-successful dog) or a False Negative (Fail label incorrectly given to successful dog). A useful model would consist of large numbers of True Positives and True Negatives, resulting in high *true positive rates* (*tpr*) and high *true negative rates* (*tnr*) (Fawcett, 2006). The *tpr* of our results was 0.93, indicating a 93% chance that a successful dog was correctly labeled with a Pass outcome. However the *tnr* was 0.07, indicating only a 7% chance that a non-successful dog was correctly labeled with a Fail outcome.

Table 2.5. Matrix of predicted vs. actual dog outcomes.

		Actual Outcomes	
		Non-successful	**Successful**
Predicted Outcomes	**Fail**	2	5
	Pass	27	70

Notes. Upper left quadrant is number of true negatives. Upper right quadrant is number of false negatives. Lower left quadrant is number of false positives. Lower right quadrant is number of true positives.

The *tpr* and *tnr* do not take outcome distribution into account. Our logistic model contains more information about successful dogs than non-successful dogs because successful dogs make up a larger proportion of our sample. Therefore it is not surprising that our model would predict Pass probabilities over the 0.50 threshold at a high rate. It is

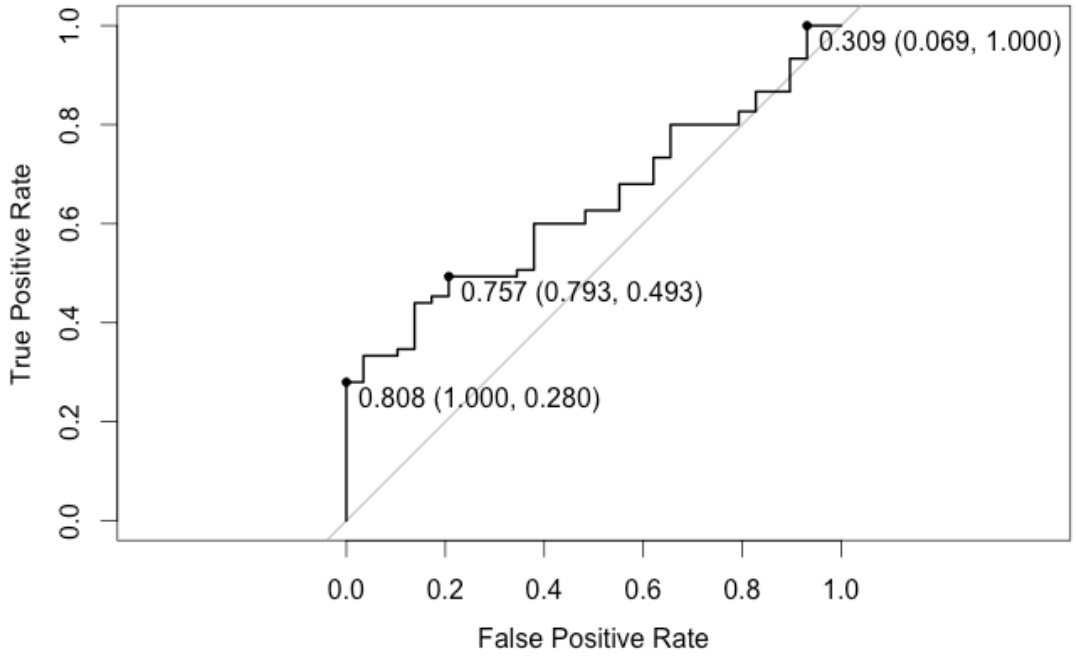
possible that increasing this 0.50 threshold, thereby making the criteria for classifying a dog as a Pass outcome stricter, would give the model the opportunity to more accurately predict non-successful dogs (Fawcett, 2006). For example, if only predicted probabilities exceeding the 0.60 threshold were labeled as Pass outcomes, then cases with probabilities lying between 0.50 and 0.60 would be labeled with Fail outcomes, where they had previously been labeled as Pass outcomes. Using stricter thresholds in this way would likely increase the model's ability to correctly label Fail outcomes (increasing *tnr*), but only at the expense of decreasing its ability to correctly label Pass outcomes (decreasing *tpr*). Therefore, it is important to find a threshold that optimizes this tradeoff point, correctly predicting as many Fail outcomes while still correctly predicting Pass outcomes at as high a rate as possible.

To measure the costs of the tradeoffs at various thresholds, we constructed an *ROC curve* (Figure 2.5). The *tpr* is plotted on the y-axis. Subtracting the *tnr* from 1 gives us the proportion of false positives (*false positive rate* or *fpr*) and is plotted on the x-axis. The line starting at the 0,0 coordinates and running 45 degrees diagonally to the 1,1 coordinate indicates the points in the graph where predictions made are no better than 50% accurate (i.e., no better than random chance). Points on the ROC curve are plotted at various thresholds up the y-axis. Points to the left of the diagonal indicate better than chance predictions. Points along the diagonal line indicate no better than chance predictions. Points will rarely lie to the right of the line, as at the very least, a predictive model should be able to predict equally well as random chance. A perfectly predicted model would show a straight line running vertically from 0,0 to 0,1, and horizontally from 0,1 to 1,1.

The first point on the ROC curve in Figure 2.5, closest to the x-axis, shows that when the threshold for receiving a Pass label is .80, the *fpr* is 0.0. This means that when a

Pass label was made only when the threshold was over 0.80, all non-successful dog outcomes were correctly labeled as fail (*tnr* = 1.00), but only 28% of successful dogs were predicted correctly (*tpr* = 0.28). Thresholds decrease at each point moving up the y axis until the final point at the top end of the y axis, where the Pass threshold is 0.32 and the model is able to correctly label all successful dogs (*tpr* = 1.0), but only 7% of the non-successful dogs (*tnr* = 0.07), which is no better than our *tnr* at the original 0.50 threshold. The best threshold in terms of being able to most accurately classify both successful and non-successful is 0.76. At this threshold, the model accurately assigns Fail labels to non-successful dogs 86% of the time (*tnr* = 0.86). However, this comes at the high cost of only accurately assigning Pass labels to successful dogs 44% of the time (*tpr* = 0.44).

The area under the ROC curve (AUC*AUC*) statistic reduces the ROC curve to a single digit ranging between 0.00 and 1.00 for measuring overall performance. The *AUCAUC* statistic represents the probability that a randomly chosen true positive case (successful dog) will have obtained a higher probability outcome prediction than a randomly chosen true negative case (non-successful dog). An AUC*AUC* of 0.50 would indicate that a model is performing at random chance, and an AUC*AUC* of 1.0 would indicate perfect prediction. Our results show an AUC*AUC* of 0.64 (L95% C.I. = 0.53, U95% C.I. = 0.75). (*CI* = 0.53-0.75), indicating a probability of just above chance that our model would predict a higher probability of a Pass outcome for a randomly chosen successful dog than for a randomly chosen non-successful dog. Ultimately, the PRS Survey was unable to make predictions about non-successful dogs accurately.



Notes. Thresholds shown in following format: “threshold level (tpr, tnr)”. Nearest to x-axis shows threshold for perfect tpr. Furthest from x-axis shows threshold for perfect tnr. Threshold for best trade-off is 0.757, where the tpr is 0.793 and tnr is 0.493.

Figure 2.5. ROC curve. Identified points at three threshold levels are shown.

DISCUSSION

Puppies bred and raised in detector-dog programs have traditionally been assessed and tested using behavior ratings. This procedure may limit program capacity to capture relevant behavior for ultimate working success because behavior ratings are often restricted in their measurement breadth (i.e., they measure a narrower range of observed behavior compared to subjective ratings). Subjective ratings may enhance the AC&BPS's ability to capture a broader range of observed puppy behavior thereby enabling stronger prediction of ultimate working success at earlier ages. We developed a new 27-item SR ratings instrument designed to cover all relevant adult detector-dog behavioral traits for use in the AC&BPS detector dog program; we then deployed our questionnaire survey to Australian households raising puppies for the AC&BPS.

We conducted a PCA using all single trait survey items (omitting an overall evaluation item) to determine whether the 26 traits measured broader underlying dimensions. Of the seven factors extracted from the PCA process, only Trainability was able to demonstrate predictive ability of outcomes. This finding is consistent with past research that suggests dog traits similar to our Trainability factor may be important for successful working dog outcomes (Maejima et al., 2007; Rooney, Gaines, Bradshaw, & Penman, 2007). A cross validation analysis assigned each dog a Pass/Fail label using Trainability to predict the dog's probability of obtaining a Pass outcome. Predicted probabilities over 0.50 were given Pass labels and probabilities under 0.50 given Fail labels. To determine how accurately the model was able to predict real life outcomes, the generated Pass/Fail labels were compared to the real life outcomes of each dog. Our

findings indicated that Trainability excessively predicted probabilities over the 0.50 threshold, resulting in the over-assignment of Pass labels. The high Pass prediction rate is not surprising when we examine the effects of obtained Trainability scores on predicted probabilities. At the 0.50 threshold, the model predicted that any dog scoring over a 4.14 in Trainability would have a successful real world outcome. Only seven dogs in our sample (7%) received scores under 4.14. This means that dogs were not predicted to fail very often. Furthermore, when dogs were predicted to fail, the model was usually incorrect, with five of the seven dogs predicted to fail actually going on to real world successful outcomes. Overall, Trainability scores were high for the dogs in our sample. Therefore, we used ROC curves to increase threshold points, so that dogs would need to have received higher Trainability scores for our model to make pass predictions. For example, perhaps if dogs needed Trainability scores over 6 to be given pass predictions, then there would be more opportunity for the model to accurately predict non-success. However, our results showed that there was no threshold at which the model could predict non-success better than random chance.

Puppy raisers generally assigned their dogs high ratings for positive traits and low ratings for negative traits, resulting in little overall variability in survey responses relative to mean responses for individual items. In some cases (for "aggression" items and "playfulness") an extreme lack of variation across survey participants was observed. The lack of variation is most likely the result of artificial breeding selection for particular behaviors in the AC&BPS program; their breeding protocol appears to have resulted in a population of dogs (relative to others, such as a pet-dog population) with very similar

characteristics that are desirable for detection work. Puppy raisers may therefore be rating the traits of detector dogs highly in comparison to companion pet animals. It is possible that to make ratings that are more accurate, puppy raisers need more definitive guidelines and examples of behaviors that differentiate how dogs should be rated across the scales of each trait. It is also a reasonable possibility that participants assigned high ratings as a result of response bias for the dogs they had been raising through puppyhood (Podberscek & Gosling, 2000). If puppy raisers are acting as de facto owners as we suggest, it is not surprising that most would exhibit positive feelings towards their dogs' prospects.

We ran two follow-up regression analyses testing the relationship between Trainability scores and dog age. For the first analysis, we included all dogs and found a significant, but small effect of age on Trainability ($b = 0.01$, $p < 0.01$, $R^2 = 0.07$). We removed 31 dogs above 15 months old from the second analysis because it was likely their outcomes had already been determined, possibly influencing the Trainability scores and the high positive skew of dog age could be causing the small effect size. Removing dogs older than 15 months resulted in a stronger effect size of age on Trainability scores ($b = 0.10$, $p < 0.01$, $R^2 = 0.11$). These results suggest that in general, puppy raisers view candidate dogs as more trainable when they are older, than when they are younger. Interestingly, of the seven dogs predicted to Fail using the 0.50 threshold as the Pass criteria, five came from only two different litters. Both of these litters were born in the same month (December, 2011), making the puppies between 5-6 months at the time surveys were given. These five puppies make up 55% of all 5-6 month old puppies

included in our study. The age range of these puppies roughly corresponds with the *juvenile period* of puppy development, during which puppies begin to sexually mature (Scott, 1958), motor skills develop, and dogs begin exploring the environment independently (Scott, 1965). Dogs will also often display increased likelihoods of exhibiting fearful reactions when encountering novel stimuli during the juvenile period (J. Serpell & Jagoe, 1995). Companion dog trainers often refer to 5 to 6 month old period as a time when dogs' manners deteriorate and they stop responding to the owners' requests (e.g., Dunbar, 2001). Therefore, it is quite possible that some raters were experiencing increased misbehavior from 5-6 month old puppies, resulting in lower ratings on the Trainability items included on our survey.

An inter-rater reliability analysis indicated that human raters thought that most survey items were generally applicable to a detector dog population. We obtained strong inter-rater reliability estimates for 13 of the PRS Survey items. The item exhibiting the strongest reliability ($\partial = 0.81$) was the overall evaluation item "Do you think this dog will make a good detector dog?", indicating strong agreement between raters on the overall potential of the dogs they were rating. The overall evaluation item was examined for predictive ability separately from the single-trait items. Most raters scored their dogs highly on this item ($M = 5.95$, $SD = 1.17$), suggesting that as a whole, raters thought highly of the potential of successful outcomes for their dogs. If raters were accurate, high evaluation scores would be expected, because most of the dogs in our sample went on to successful outcomes. However, the results of a logistic regression model evaluating the overall evaluation item as a predictor of dog outcome suggest that raters were not

accurate in their assessments. Therefore, the high ratings in the overall evaluation item could again be a result of the narrow range of the detector dog population. Bias did not impact puppy raisers' abilities to predict outcomes using a similar assessment item in a population of guide dog candidates (Batt et al., 2009), but that possibility is present for this study (Podberscek & Gosling, 2000). The strong reliability of the broad evaluation item indicates that although their predictions were inaccurate, multiple members of households generally agreed on the overall potential of the dog they were rating.

Six of the 12 single trait items showing strong inter-rater reliability loaded onto the Trainability factor extracted in our PCA. Of the remaining six factors extracted in our PCA, only Energy Level (which consisted of only two items) and Emotional Arousal (our weakest factor) consisted of more than one item with high inter-rater reliability. Additionally, Trainability was the only factor to demonstrate a significant correlation with outcome. These patterns suggest that the Trainability items were better at capturing relevant data from puppy raisers than were the items pertaining to the other six factors. However, past research suggests that personality dimensions similar to some of these factors, such as Confidence and Energy Level, may be important to detector-dog success (Maejima et al., 2007; Svartberg, 2002). Therefore, developing survey items that more reliably measure non-Trainability traits could lead to stronger factors that when combined with Trainability might improve the predictive model.

We ran a second PCA analysis dropping the five traits that had shown weak inter-rater reliability estimates. The resulting seven-factor structure was slightly different than the structure using all items, however an LOOCV analysis showed the same Trainability

factor to be the only significant predictor of outcome when the weak inter-rater reliability items were removed.

Two items, “independent” and “kennel behavior”, resulted in large proportions of N/A (non-applicable) ratings. The definition used to describe the item “independent” referred to the dog’s ability to work alone. We suspect that many of the puppy raisers related the word “work” to detection work, which they would not have the opportunity to witness. For the item “kennel behavior”, we determined after initiating data collection that the AC&BPS does not permit the raisers to kennel their puppies; this may explain why close to 20% of participants chose to rate this trait as ‘N/A’ in their AC&BPS puppy.

LIMITATIONS AND FUTURE DIRECTIONS

All research efforts have limitations, and we note three limitations below. We also make recommendations for future studies that may help address these limitations.

One limitation of this research is that successful dogs outnumbered unsuccessful dogs in our sample by nearly three to one. The unbalanced sample means that the predictive model had less information about non-successful dogs than it did for successful dogs. The disproportion makes it difficult to render solid conclusions as to why the model was unable to predict non-successful outcomes. Collecting additional samples would help accumulate more information about non-successful dogs. It also may help to change the criteria for determining success so that more dogs are assigned Fail labels. Studying dog populations that have more evenly balanced outcome rates is another potential solution.

A second, related, limitation is that the AC&BPS dog population consists of dogs that are purpose-bred to possess specific characteristics desirable for detection work. Therefore, it is possible that many dogs that are ultimately disqualified from the program still exhibit these characteristics at moderate levels, when compared to the general pet population. If true, then puppy raisers may be rating detector dogs in comparison to the pet dog population, resulting in disproportionately high ratings. For example, a Labrador Retriever puppy the AC&BPS might consider low in toy drive might be considered high in toy drive by the puppy raiser compared to his/her pet Chihuahua. Future surveys may need to ensure that puppy raisers use the entire range of the trait scales. Item descriptions may need to more precisely define the high and low presence of traits specific to the

detector-dog population to help raters discriminate between finer levels of the traits. Including more examples of how certain behaviors could manifest themselves in situations relevant to the puppy raiser environment may help add clarity to these items. For example, "playfulness" did not receive strong agreement so revisions of item definitions to include initiating and maintaining self-play may be more generalizable across observers. Photographs or videos of example behaviors may also be useful for puppy raisers.

A third limitation is that outcomes were broadly categorized as successful and non-successful. However, the AC&BPS will sometimes sell dogs to organizations such as the police force or to international working dog programs. No distinctions were made based on these types of working outcomes because AC&BPS considered all dogs placed with any organization were considered successful and additionally, only nine dogs from our sample became active AC&BPS non-breeding detector dogs. One reason that AC&BPS supplies dogs to other organizations is likely to be a matter of numbers. Less than 70 dogs actively work as detector dogs for the AC&BPS ("Detector Dog Program - Working with Dogs to Help Protect Australia," 2012). Yet the sample we obtained over a very limited time period yielded 74 dogs that went on to successful outcomes (66, when not including breeder dogs). It is unlikely that there would have been enough available AC&BPS detector dog positions for all 74 of these dogs. So the question of why dogs are selected for AC&BPS as opposed to being sold to other organizations arises. All the dogs would have shown characteristics that make them suitable for working outcomes but it is possible that behavioral "tiers" are considered when the AC&BPS chooses whether a dog

should be kept or sold to another organization. The current study considers all successful dogs to be behaviorally equal, and therefore does not account for any such tiers if they exist. Future studies should attempt to account for any such differences in success tiers.

CONCLUSION

Measuring broad personality traits through subjective methods can be particularly useful because it can allow for relevant behaviors to be measured in a reliable way; therefore, subjective information could be used to select dogs with high likelihood of success, to efficiently track dog behavior through time, and to design cost-effective developmental interventions. This research shows that puppy raisers may be able to contribute information about candidate dogs that is helpful for AC&BPS using subjective ratings surveys. Strong reliability was found in several survey items making up a factor called Trainability. Trainability was able to predict a large portion of successful outcomes. However, our predictive model would not be useable in its current state because it was prone to over-predicting success and poor at predicting non-success in our sample. Possible reasons for these results and suggestions for improving the survey were identified and discussed.

Chapter 3

The Predictive Abilities of Personality, Attitudes towards the Human-Dog Relationship, and Sports Participation On Detector-Dog Handler Job Performance

BACKGROUND

In Chapter 2, we examined dog traits that past research and expert opinion have identified as beneficial for detector dog success. Now, in Chapter 3, we explore whether certain human traits might be beneficial for the success of the human detector-dog handler. In particular, we examined the extent to which a questionnaire measuring demographics, personality, attitudes towards the dog-human relationship, and sports experience could predict job performance in AC&BPS detector-dog handlers.

Demographics and Personality in Job Personnel Performance

In the field of human job personnel research, subjective supervisor ratings are frequently used as measures of employee performance (Sturman, 2003). Supervisor ratings of job performance have been examined in relation to various employee demographics such as age, job seniority, and gender. Empirical studies of the relationship between age and job performance have shown mixed results in several fields, including managerial, engineering, supervisory, clerical, sales, and bank teller jobs (McEvoy & Cascio, 1989; Ng & Feldman, 2008; Sturman, 2003), with some meta-analyses failing to find consistent support for a relationship between age and supervisor ratings (McEvoy & Cascio, 1989) and others finding an effect (e.g., Sturman [2003] who found an effect up until the age of 50). It is sensible to think that age would positively correlate with job performance because experience often comes with age and as employees gain more

experience in their workplace, they should become better at their jobs. Over time, many of these more experienced employees are likely to be promoted to senior-level positions. Therefore, if age is positively associated with job performance, then job seniority should also positively associate with job performance. As a result, if age is associated with better job performance, then job seniority should also correlate. On the other hand, moderating factors that differ between job types, such as job complexity, physicality, and skill types, may cause age and job seniority to have the reverse effects. For example, as employees get older, physical abilities may diminish, resulting in inferior job performance in highly physical jobs (Sturman, 2003). It may be that age is more predictive of more specific job related dimensions, rather than overall performance. For example, older workers have been found to be more likely to abide to and be supportive of organizational norms, and less likely to engage in "counterproductive work behaviors" (Ng & Feldman, 2008).

In certain contexts, employee gender may contribute to performance ratings given by supervisors. For example, gender effects have been demonstrated when a job is stereotypically "male" or "female" (see Davison & Burke, [2000] for a review). Females were rated as inferior at car and heavy machinery sales, jobs that were categorized as "male sex- typed." Meanwhile, males were rated as inferior at "female sex- typed" jobs, such as secretarial work and home economics teaching (Davison & Burke, 2000).

In the field of human personality assessment, research has shown that certain personality traits predict future job performance across a variety of professions, including management, sales, and police work (Barrick & Mount, 1996; Barrick, Mount, & Judge, 2001; Penney, David, & Witt, 2011). Personality traits can be thought of as enduring behavioral, emotional, and mental characteristics that differ between individuals (Goldberg, 1999). With the widespread acceptance of a standard personality taxonomy—the so-called Big Five (John, Naumann, & Soto, 2008) — personality testing has become

an increasingly common way for professional organizations to prescreen job applicants (Paul, 2010).

The Big Five consists of five broad personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) that are thought to capture the bulk of variance in human personality traits (John et al., 2008). Different characteristics are commonly associated with the high and low poles of each bipolar dimension. High Openness is typically associated with creativity, open-mindedness and originality, while low Openness is linked with relatively traditional and conventional attitudes and behaviors. High Conscientiousness is commonly defined by goal directed activity, impulse control and dependability, while low Conscientiousness is associated with carelessness and lack of impulse control. High Extraversion is typically associated with sociability, high activity, and enthusiasm, while low Extraversion is associated with less assertive, quieter behavior. High Agreeableness is commonly associated with trustworthiness and an ability to work well with others, while low Agreeableness tends to be linked with antagonism and selfishness. High Neuroticism is associated with anxiety, nervousness, and personal insecurity, while low Neuroticism is defined by even-temperedness and emotional stability (Barrick et al., 2001; John et al., 2008).

Of the five dimensions, Conscientiousness has proved the most effective at predicting job performance across occupational domains, and Neuroticism has shown predictive value among police and skilled/semi-skilled occupations (Barrick et al., 2001; Barrick & Mount, 1991). It is not surprising that goal directed, dependable (high Conscientiousness) employees do better jobs than careless, impulsive (low Conscientiousness) employees, regardless of the type of profession (Barrick et al., 2001; Barrick & Mount, 1991). Nor is it surprising that emotionally stable, even-tempered (low Neuroticism) employees in many professions tend to do better jobs than anxious,

nervous, and insecure (high Neuroticism) employees (Barrick et al., 2001; Barrick & Mount, 1991). The remaining three dimensions, Openness, Extraversion, and Agreeableness, have not been shown to consistently predict job performance (Barrick et al., 2001; Barrick & Mount, 1991).

In job personnel studies using personality measures, it is plausible that employees may purposely answer survey items dishonestly to appear more favorably to employers. This form of impression management could distort Big Five personality scores for some participants (Paunonen & LeBel, 2012). Big Five measures have shown to be generally resistant to effects of dishonest answering (Ones & Viswesvaran, 1998) but highly distorted traits can cause problematic results when interpreting the results of measures at individual levels (Paunonen & LeBel, 2012). Therefore, in contexts where participants are motivated to convey a positive impression it is prudent to assess socially desirable responding.

Human Personality and Dog Behavior

Human personality traits may be associated with handler job performance through their effects on dog behavior. It is possible that dogs may behave differently or perform less effectively for humans possessing certain personality traits. For example, companion pet dogs belonging to owners with low Neuroticism can demonstrate fewer stress behaviors compared to dogs belonging to owners with high Neuroticism when performing a series of operational tasks, such as learning simple tricks and being led over a mesh bridge (Kotrschal, Schöberl, Bauer, Thibeaut, & Wedl, 2009). Additionally, task performance was better for dogs whose owners were high in Agreeableness and Conscientiousness (Kotrschal et al., 2009).

Dog-human Relationships and Dog Behavior

The dog-human relationship has received increasing attention recently, particularly among the pet dog and owner population (e.g., Bennett & Rohlf, 2007; Hsu & Serpell, 2003; Rooney & Bradshaw, 2002; Topál et al., 1997). The quality of the relationship appears to be an important factor in determining dog behavior (Kotrschal et al., 2009; Topál et al., 1997). Past research has mostly used the amount of obedience training, time spent together, and level of attachment as proxies of relationship quality and the presence or absence of dog behavior problems as an outcome measure (Jagoe & Serpell, 1996; Kobelt, Hemsworth, Barnett, & Coleman, 2003). Evidence suggests that pet dogs exhibit fewer behavior problems for owners who report stronger attachments and spend more time with them (Bennett & Rohlf, 2007; J. a. Serpell, 1996).

However, these studies all target companion pet dogs. The nature of the handler-working dog relationship may be very different, and may vary according to the type of working role. Additionally, the absence of behavior problems may not be diagnostic of a dog's ability to perform a specific function well. Unfortunately, research into the human-perceived handler-working dog relationship has been scarce. sent surveys to multiple military dog organizations questioning handlers about the nature of their relationships with their assigned working dogs. Results showed that the dogs were more obedient for handlers that spent more time with them outside of the workplace, actively strengthening the relationship through shared activities like canine sports. Presumably, handlers who actively worked to strengthen the relationship did so in part because they believed it led to better performance. This finding suggests that handlers who put greater value into building strong relationships and bonds outside of work may achieve better performance results in the workplace from their detector dogs.

Participation in Sports as a Predictor of Dog Handler Performance

In addition to using previous research to generate candidate predictors of handler performance, we also surveyed the intuitions of the experts working at AC&BPS. We reasoned that after many years observing successful and unsuccessful handlers pass through the program, seasoned veterans probably would have acquired some insight into what human characteristics predict dog-handler performance. Our panel of odor-detector dog experts suggested that prior participation in sporting activities, particularly when participating at higher skill levels, might contribute to better job performance of handlers in the future. The AC &BPS experts believed that sports and detector-dog handling required similar skill sets: hand-eye coordination, good snap-decision making, and physical execution in high-pressure environments. There is evidence within the sports-performance literature that is consistent with these intuitions. For example, it has been found that expert athletes of a single sporting activity apply decision and pattern recognition skills previously acquired through participation in a wide range of multiple alternate sports (Baker, Cote, & Abernethy, 2003; Bridge & Toms, 2013). Therefore, it is possible that decision and pattern recognition skills learned through sporting activities are later incorporated into the jobs of detector-dog handlers.

HYPOTHESES

Based on past research and the expert intuition discussed above, we developed four hypotheses regarding predictors of handler success.

Hypothesis 1: Demographics

We predicted that job seniority would correlate with better job performance in AC&BPS dog handlers, independent of gender. In the current study, we predicted that age would predict better performance because of likely correlations with higher seniority

levels. We did not make a prediction based on gender because of the disproportionate amount of male ($N = 39$) vs. female handlers ($N = 9$) in our sample.

Hypothesis 2: Personality

Given a complete lack of published information on personality and working-dog handler performance, we used past research examining job performance spanning multiple professions as the basis for our hypothesis. We predicted that high Conscientiousness and low Neuroticism scores would predict high job performance in AC&BPS dog handlers (Barrick et al., 2001; Barrick & Mount, 1991), independent of age, job seniority, gender, and impression management. We did not make predictions for Openness, Extraversion, and Agreeableness because of the lack of findings in previous research (Barrick et al., 2001; Barrick & Mount, 1991). However, the specialized field of working-dog handling has not yet been evaluated by personality research, and we therefore cannot assume that Openness, Extraversion, and Agreeableness do not contribute to the success of working dog handlers, so they have been included in our analysis.

Hypothesis 3: Attitudes towards the dog-human relationship

We predicted that handler attitudes strongly valuing the dog-human relationship would correlate with better job performance than attitudes not valuing the relationship, independent of age, job seniority, and gender.

Hypothesis 4: Sports Participation

We predicted that playing and/or being skilled at a wider variety of sports would correlate with better job performance in AC&BPS dog handlers than playing and/or being skilled in fewer sports, independent of age, gender, and job seniority.

METHODS

Participants

Sixty-three officers employed by the AC&BPS detector-dog program were asked to complete a two-section self-report questionnaire. The first section consisted of items to capture participants' demographics, attitudes towards the detector dog-handler relationship, and experience playing sports. The second section consisted of an impression management scale, a personality scale, and a scale measuring attitudes towards pets. Handlers were notified about the questionnaire on July 2, 2012. The final completed questionnaire was received on November 8, 2012.

AC&BPS officers held one of three job seniority levels. The entry-level position of *level 1 handler* performs search tasks at locations such as airports, seaports, and postal facilities as one half of a dog-handler team. The higher-ranked *level 2 team leaders* perform searches with detector dogs as well as oversee small groups of 4-5 dog-handler teams. The term *handler* is used in the present study to broadly refer to all officers independent of level 1 or level 2 statuses. *Instructor/supervisor*s were the highest-ranked officers to participate in this study. Instructors/supervisors select and develop both dogs and handlers for suitability in detection work, and manage large groups of 7 to 16 level 1 and level 2 handler-dog teams.

Fifty-two officers returned questionnaires, representing a return rate of 83%. Four of the 52 (8%) questionnaires were submitted by instructor/supervisors. Linear predictive models could be impacted by an unbalanced design in which one of our comparison groups only featured four members. Additionally, these instructor/supervisors were

responsible for assigning the job performance ratings to all AC&BPS handlers, including themselves (see below). Therefore, we excluded the level of instructor/supervisor from our analyses. This resulted in a sample of 48 participants. Of the 48 cases, 38 (79%) completed both sections of the questionnaire. Ten participants completed only the first section of the questionnaire. As a result, sample sizes for predictive analysis differed between predictive models, depending on what variables were in the analysis.

The majority of the 48-case sample was male (81%). One male handler did not indicate his age. The mean age of remaining males was 35.32 (*SD* = 4.20) and the mean age of females was 32.44 (*SD* = 3.01). Eighty-three percent of handlers and 77% of team leaders were male.

Measures

Job Performance Ratings

Job performance was measured using a competency score scale that ranged from 1 (lowest competency level) to 7 (highest competency level), with a 4 indicating average competency relative to the other potential participants.

Demographics

Participants were asked to indicate their age, job position (level 1 handler, level 2 team leader, or instructor/supervisor), and gender.

Personality

To measure the Big Five Personality dimensions, we used the 120- item IPIP-NEO scale (Appendix F) that yielded scores on participants' levels of Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The IPIP-NEO is derived from the International Personality Item Pool (IPIP), a larger set of personality measures developed and tested for public domain use by the scientific community (Goldberg et al., 2006). Items on the IPIP-NEO consist of personal characteristics on which the participants rate themselves on a scale of 1 (*Very Inaccurate*) to 5 (*Very Accurate*). The language of two politically related items (items 28 and 88) was changed to account for a difference in lexicon for the Australian population. A single score for each Big Five personality dimension was obtained by averaging the scores of all the IPIP-NEO items that represented the dimension. Thus, each participant had five unique scores ranging from 1 to 5, each of which captured the participant's measurement position between the two poles of a Big Five dimension.

The context of the study, in which participants were being evaluated at work, raised the possibility that participants might answer personality items dishonestly in an attempt to appear more favorable, potentially resulting in distorted personality scores (Paunonen & LeBel, 2012). Therefore, we administered the 20-item Impression Management (IM; Appendix G) portion of the Balanced Inventory of Desired Responding (BIDR; Paulhus, 1991), which is designed to measure purposeful impression management. Each item on the IM scale is answered using a 7-point scale ranging from 1 (*not true*) to 7 (*very true*). IM scores are calculated by awarding one point for each

extreme answer (a 6 or a 7; negatively worded items are re-coded to positive scales). The total number of extreme answers is the participant's IM score; higher scores indicate that the participant was more likely to be distorting personality questionnaire item answers to appear more favorably, lower scores indicate that the participant was less likely to be distorting answers.

Attitudes towards the dog-human relationship

To our knowledge, there are no validated instruments for measuring dog-handler attitudes towards their relationship with working dogs. However, there are validated measures of human attitudes towards their companion pet dogs; the majority of participants in our sample currently owned at least one pet dog (62.5%) and 77% had previously owned a dog. Therefore, we used the Pet Attitude Scale (PAS) developed by Templer (1981) and later slightly modified (PAS-M; Appendix H; Munsell, Canfield, Templer, Tangan, & Arikawa, 2004), to measure handler attitudes towards pets. The PAS-M is the most frequently used pet attitude measurement and has shown previous evidence of reliability and validity (Preylo & Arikawa, 2008). The PAS-M contains 18 items that the participant scores on a scale of 1-7. These items have been shown to collectively represent a single underlying index of negative vs. positive attitudes towards pets (Preylo & Arikawa, 2008), with higher scores indicating more positive attitudes. PAS-M items are phrased in a fashion that allows subjects to answer the items even if they do not own any pets. A single pet attitude score was derived for each participant by averaging the scores of the 18 PAS-M scale items.

We added two items to the survey via consultation with AC&BPS supervisors (Appendix I) that aimed to measure the handlers' attitudes towards the human-dog

relationship within the specific detector dog/handler team context. An "emotional bond" item asked participants to rate how important they felt it was for handlers to maintain strong emotional bonds with their detector dogs. A 1-5 scale (1 = *not important at all*, 3 = *neither important nor unimportant*, 5 = *extremely important*) was used for the emotional bond item. A "relative importance" item asked the participant to rate how important the quality of the dog vs. the quality of the handler was for the success of dog/handler team. A 1-5 scale was used for the relative importance item (1 = *the quality of the dog is the only important part*, 3 = *the quality of the dog and the quality of the handler are equally important*, 5 = *the quality of the handler is the only important part*).

Sports Participation

Questionnaires included a list of twenty sports assembled in conjunction with AC&BPS supervisors as a measure of sports participation (Appendix J). Participants were asked to click a checkbox adjacent to each sport to indicate having played that sport. The number of checkboxes selected was added up for each participant and converted into a variable representing the total number of sports played (total sports). The largest possible total sports value was 20. The total sports measure did not capture any information about the frequency of activity or ability level attained at each sport. Therefore, we included additional items to distinguish between "competition-level" and "social-level" play, speculating that competition-level play would indicate higher ability level and stronger involvement. When participants indicated experience in a particular sport, they were then presented with additional items asking whether the sport was played at a social-level or at competition-level. From these results, a competition-level sports variable was created that included only the total number of sports played at competition-levels.

Procedure

Potential participants were sent an email invitation and study description from an AC&BPS supervisor. The email gave assurances that identities would be kept confidential and that participation would have no effect on job status. A link to an online informed consent form was included in the email. After providing informed consent online, participants were taken to the first part of the online self-report questionnaire. Upon submitting the first part, they were immediately taken to the second part of the questionnaire. Participants were able to leave the questionnaire at any time, save their responses up to that point, and return to where they had left off later using a password.

To obtain competency scores, a single supervisor/instructor stationed at the main AC&BPS headquarters was sent an email link to an online form listing all potential participants. Level 1 handlers and level 2 team leaders were deployed throughout Australia working under different supervisors/instructors. As a result, some supervisors were familiar with the job performance of only some handler/team leaders. Therefore, competency scores were determined through internal discussion and consensus among supervisors familiar with each particular participant (i.e., not all participants were given competency scores by the same set of supervisors). To encourage the supervisors to use the entire range of the scale, they were first asked to score all participants they considered 1's (lowest competency) and 7's (highest competency). They were then asked to rate the remaining participants in relation to those extremes. Supervisors were assured that competency scores would be kept confidential from handlers and team leaders. Handlers and team leaders were not made aware that they would be receiving competency scores at the time they were asked to complete their questionnaires.

Statistical Analysis

Fourteen predictor variables were included in our statistical analysis. These consisted of three demographic variables, five personality scores, one impression management score, three attitudes towards the dog-human relationship scores, and two sports variables. Descriptive statistics and Pearson correlation coefficients were computed using all pairwise combinations of the 14 variables (Table 3.1). Ten participants failed to complete the second section of the questionnaire, which consisted of the impression management, personality, and pet attitude scales. This resulted in sample sizes differing across pairwise comparisons, which are also given in Table 3.1.

Table 3.1. Descriptive Values and Pearson Coefficients of variables measuring Demographics (DEM), Personality (PERS), Attitudes towards the human-dog relationship (ATT), Involvement in Sports, (ALL-SPRT), and Involvement in Competitive-level Sports (CL-SPRT). Diagonals indicate alphas of scales.

Category	Variable	*N*	*M (SD)*	Age	Job Pos.	Gender	IM	O	C	E	A	N	Pet. Att.	Bond	Rel. Imp.	Tot. Sp.
DEM	Age	47	34.77 (4.13)	——												
	Job Pos.	48	0.27 (0.45)	0.31	——											
	Gender	48	0.81 (0.39)	0.28	-0.07	——										
PERS	IM	38	10.34 (4.46)	-0.10	-0.08	-0.12	0.89									
	O	38	3.09 (0.40)	-0.03	0.05	0.15	-0.14	0.79								
	C	38	4.06 (0.49)	-0.04	0.15	-0.3	0.62	-0.19	0.89							
	E	38	3.56 (0.47)	-0.16	-0.08	0.05	0.14	0.12	0.35	0.80						
	A	38	3.91 (0.40)	-0.24	0.02	-0.35	0.41	0.10	0.39	0.17	0.86					
	N	38	2.20 (0.58)	0.08	0.19	0.03	-0.47	-0.01	-0.56	-0.51	-0.41	0.92				
	Pet. Att.	38	5.20 (0.67)	-0.34	0.11	-0.21	0.26	0.22	0.37	0.38	0.13	-0.04	0.39			
ATT	Bond	48	3.62 (1.06)	-0.49	-0.09	-0.17	-0.52	0.18	-0.37	0.04	-0.15	0.17	0.04	——		
	Rel. Imp.	47	2.96 (0.20)	0.08	0.13	-0.10	-0.15	0.07	0.00	-0.09	-0.16	0.01	-0.11	-0.17	——	
ALL-SPRT	Tot. Sp.	48	5.48 (4.08)	0.26	0.03	0.15	-0.36	0.21	-0.32	-0.24	-0.24	0.21	-0.27	-0.10	0.10	——
CL-SPRT	CL. Sp.	48	2.56 (2.31)	0.22	-0.01	-0.05	-0.24	-0.03	-0.34	-0.15	-0.12	0.22	-0.27	-0.12	0.20	0.77

Notes: Job Pos. = Job Position, IM = Impression Management, O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism, Pet Att. = Pet Attitudes, Bond = Emotional Bond, Rel. Imp. = Relative Importance, Tot. Sp. = Total number of sports played at any level, CL. Sp. = Total number of sports played at competition level, for Gender 0 = Female, 1 = Male, for Job Position 0 = Handler, 1 = Team Leader

The total amount of missing values on the IPIP-NEO questionnaire was less than 1.0%. We did not impute missing values prior to calculating Big Five dimension scores because multiple items measured each dimension, allowing us to obtain aggregate scores for each personality trait based on the subset of items participants actually did complete (Goldberg, 1999).

The total number of missing values on the IM questionnaire was three, representing less than 1.0% of all responses. All three missing scores were for different items and came from different participants. We considered listwise deletion of these participants, but concluded that the data of these individuals could be important considering our relatively small sample size. Therefore, the missing scores for these items were substituted with the mean score of each participant, after converting reverse score items. None of the three replaced values were over 5, and therefore they did not contribute to the number of extreme answers when calculating the total IM score for the corresponding participant (Paulhus, 1991).

The total amount of missing values for the PAS-M scale was less than 1.20%. We did not impute values for PAS-M items containing missing data because all items measured a single pet attitude construct. Therefore, for cases missing values, the pet attitude aggregate score was computed using only the answered items.

Three sample sets were derived from groups based on the proportion of the questionnaire completed by the participants:

1. *Sample59* included a group of participants who completed sections 1 and 2 (N = 38), a group that only completed section 1 (N = 10), and a group that did not complete any portion (N = 11). Sample59 was used to test whether differences in competency scores differed between groups based on the proportion of the questionnaire completed. Two-sample t-tests were used for this procedure.

2. *Sample48* included the group that completed sections 1 and 2 ($N = 38$) and the group that completed section 1 only ($N = 10$). Therefore, we were restricted to only fitting linear regression models to Sample48 that tested our demographics and sports hypotheses. Our first model consisted of job seniority and age as predictors of competency score, and included gender as a covariate. Because one individual did not submit his age, his data was not included in this demographics model (DEM47) or any subsequent models. Separate linear models were created using variables for sports played at any level (ALL-SPRT47) and at competition-levels (CL-SPRT47). The job seniority, age, and gender items were included as covariates in these models to determine whether sports predicted competency above and beyond the effects of demographics.
3. *Sample38* included only the group that completed entire questionnaires, sections 1 and 2 ($N = 38$). Therefore, we were able to fit models to Sample38 to test all four of our hypotheses. Again, one individual was omitted because he did not submit his age. A model was created to test the effects of job seniority and age on competency score (DEM37), with gender as a covariate. A model for personality used the five personality dimension scores and impression management score as predictors of competency (PERS37). Emotional bond and pet attitude scores were used as predictors in a model to test attitudes towards the relationship on competency (ATT37). Separate models using total sports played at any level (ALL-SPRT37), and total sports played at competition-levels (CL-SPRT37) were also created. All models included the demographics variables, so that we could test the effects of our predictors above and beyond the effects of job seniority, age, and gender.

The regression coefficient (*b)* in a linear model represents the predicted change in competency score for every one-unit change of the corresponding variable, when all other variables in the equation are held constant. If the results of a regression equation show a variable's *b* to equal zero, then there is no predicted change in competency score based on that particular variable. The further a variable's *b* is from zero, the more competency score is predicted to change based on that variable. The standard error of the coefficient (*SE b*) is an estimation of the amount *b* varies across all cases. As a variable's *SE b* becomes larger, the obtained *b* becomes less trustworthy and is less likely to differ from zero. The p-value (*p)* gives the probability that there is no difference between the obtained *b* and zero. Therefore, a lower *p* means we can be more confident that a variable is having an effect on competency score that is greater than zero. The standard cut off for accepting that a result is significantly different from zero is 0.05.

Units of measurement differ across our predictor variables. For example, while job position has only two possible discrete outcomes, the variety of sports played could range from zero to 20. Therefore, we cannot use the *b* to compare the effect of a one-unit change of job position on competency score relative to the effect of a one-unit change in the sports variable. The standardized coefficients (*β)* puts all predictors on the same standardized scale, allowing us to more accurately interpret the relative importance of model predictors.

The *F* value is the statistic that tests whether the linear relationship between the entire set of predictors and competency score is significantly different than zero. The *p* for the *F* test is given for each model.

The R^2 gives the total amount of variance in competency score that is explained by each model. Larger numbers of predictors can inflate the R^2 so the *adjusted* R^2 (*adj* R^2), which penalizes models for adding predictors, is also shown. The R^2 and *adj* R^2 statistic

range from 0 to 1; closer to 1 indicates a better fitting model. The *adj* R^2 can be useful for comparing subsets of independent variables and determining a single best model, but it can not be used to compare multiple models featuring different independent variables (Whittingham, Stephens, Bradbury, & Freckleton, 2006).

Therefore, we used the information theoretic (IT) methodology developed by Burnham and Anderson (2002) to make direct comparisons of our models. The IT methodology uses the Akaike information criterion (*AIC*) to obtain a measure of fit for a model among a comparison set of competing models. Like the *adj* R^2, the *AIC* applies a penalty for each predictor as a control for the number of variables in the model. However, the *AIC* contains no measure of the global quality of a model; it only makes comparisons between the models presented. In other words, the *AIC* does not tell us whether the model is an overall good or poor fit. It only tells us how good the model fits compared to the other competing models. Thus, the *AIC* allows us to make direct comparisons between models regardless of any variation in number of predictors. A version of the *AIC*, called the *AICc,* is recommended for small sample sizes (Burnham & Anderson, 2002) and was used here to make inferences from predictive models.

To evaluate the strength of our models using the IT methodology, we created two comparison sets of models. The first comparison set (CS47) consisted of DEM47, ALL-SPRT47, and CL-SPRT47, allowing us to test our demographics and sports hypotheses using the maximum amount of data collected. The second comparison set (CS37) consisted of DEM37, PERS37, ATT37, ALL-SPRT37, and CL-SPRT37, and included only participants for which we had collected data measuring all four of our hypotheses.

CS47 and CS37 both included reference models (REF47 and REF37) containing no predictor variables. These models simply fit the mean values for competency, and test whether they differ significantly from zero. Therefore, to provide meaningful explanatory

value, predictive models would need to improve upon the reference models. *AICc* scores and weights were calculated for each model and compared to other competing models within each comparison set. The *AICc* statistic is the raw *AICc* score of each model. The *AICc* weight is the probability of each model being the best model of the comparison set.

Violations of certain assumptions can lead to inaccurate interpretations of regression results (Kutner, Nachtsheim, Neter, & Li, 2005). Therefore, assumption checks were performed for each model (Appendices K through P). Model residuals were graphed as a function of the predicted values for each model to test whether assumptions of homoscedasticity were met. We examined autocorrelations and partial autocorrelations for violations of independence. We tested normality using QQ plots to plot the model residuals against a standard normal distribution (Kutner et al., 2005). To test multicollinearity, we used the variance inflation factor (VIF). The VIF values were far below 10 for all model variables, which is generally suggested to be adequate for meeting avoiding multicollinearity problems (Kutner et al., 2005).

RESULTS

Results below are organized according to the sample set analyzed. First, we use Sample59 to show how competency scores differed according to whether participants returned Sections 1 and 2 of the questionnaire, Section 1 only, or none at all. Next, we describe characteristics of Sample48 and present results of CS47 linear models. Finally, we describe characteristics of Sample38 and present results of CS37 linear models. Table 3.2 shows all predictive models along with whether each model supported the hypotheses it was testing.

Table 3.2. List of linear models derived from obtained sample sets (Sample48, Sample38), proportion of questionnaire completed by participants in each set, hypothesized predictor variables, and whether results supported hypotheses

a. Sample48

Participants	CS 47 Models[a]	Predictor variables of competency score	Support
38 handlers completing sections 1 and 2	DEM47	Age	×
		Job position	√
10 handlers completing section 1 only	ALL-SPRT47	Total sports	•
	CL-SPRT47	Competition-level sports	√

b. Sample38

Participants	CS37 Models[a]	Predictor variables of competency score	Support
38 handlers and team leaders completing sections 1 and 2	DEM37	Age	×
		Job position	√
	PERS37	Conscientiousness	×
		Neuroticism	×
	ATT37	Emotional bond	×
	ALL-SPRT37	Total sports	×
	TOT-SPRT37	Competition-level sports	•

Notes. √ = supported ($p < 0.05$), • = marginal support ($p < 0.12$), × = not supported ($p > 0.12$)
[=] One handler from Sample38 and Sample48 was excluded from all linear models because he did not submit his age

Sample59

Competency scores ranged from 1 to 7 (M = 3.93, SD = 1.91) for Sample59 (Figure 3.1). Twenty-two participants (37%) were rated above average (competency score > 4) and 24 (41%) were rated below average (competency score < 4).

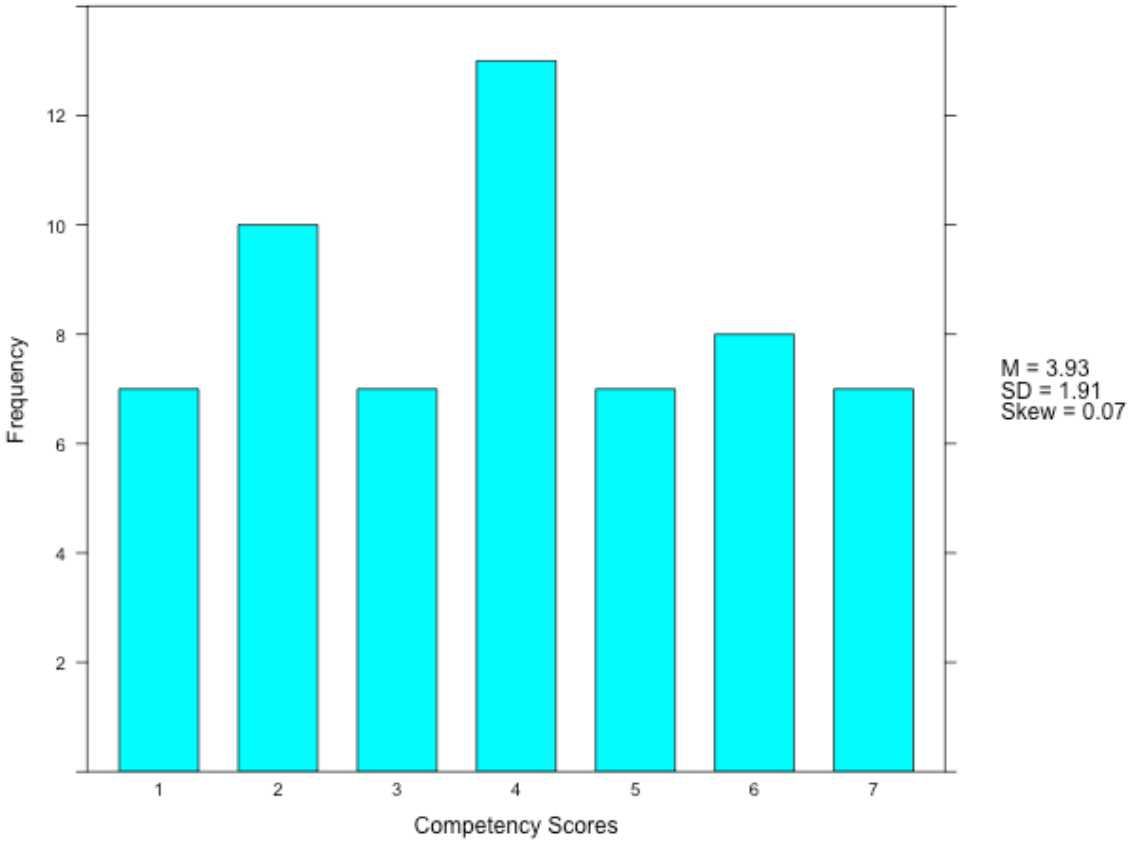


Figure 3.1. Competency score distribution of Sample59

Tables 3.3 and 3.4 show the differences in competency score in Sample59 according to the proportion of questionnaire completed. There was a significant difference ($t_{(19)} = 2.20, p < 0.05$) in competency score between the 38 participants who completed the entire questionnaire ($M = 4.55, SD = 1.88$) and the 10 participants who did not complete the personality and attitudes measures ($M = 3.40, SD = 1.30$). The Cohen's d measure of effect size indicated a $d = 0.70$, which is a moderate effect size according to Cohen's suggested rules of thumb (Cohen, 1988). There was a large effect size ($d = 1.41$) of the differences ($t_{(24)} = 4.65, p < .001$) in scores between those who completed the entire questionnaire and the 11 handlers who did not return any portion of the questionnaire ($M = 2.27, SD = 1.28$).

Table 3.3. Means and standard deviations of competency scores grouped by whether participants completed sections 1 and 2, section 1 only, or none of the questionnaire, and the percentages of all above and below average participants included in each group

Sections completed	*N*	Competency *M* (*SD*)	% of all below average	% of all above average
Sections 1 & 2	38	4.55 (1.88)	42%	90%
Section 1 only	10	3.40 (1.30)	25%	10%
No Sections	11	2.27 (1.28)	33%	0%
Total	59	3.93 (1.91)	100%	100%

Table 3.4. Distribution of handlers, team leaders, and proportion of questionnaire completed across each level of competency score

	Job Position		**Proportion of Questionnaire Completed**			
Competency Score	Handlers	Team Leaders	Sections 1 & 2	Section 1 Only	No Sections	Total
1	6	1	3	0	4	7
2	10	0	4	3	3	10
3	6	1	3	3	1	7
4	12	1	8	2	3	13
5	4	3	6	1	0	7
6	6	2	7	1	0	8
7	2	5	7	0	0	7
Total	46	13	38	10	11	59
M (*SD*)	3.52 (1.6)	5.38 (1.85)	4.55 (1.88)	3.40 (1.30)	2.27 (1.28)	3.93 (1.91)
% below average	48%	16%	27%	60%	73%	41%
% above average	26%	77%	53%	20%	0%	37%

Eight of the 11 participants (73%) who did not participate in the study (Figure 3.2) and 6 of the 10 (60%) that returned only partially completed questionnaires (Figure 3.3) were rated below average. This means that only 42% of all below average handlers returned fully completed questionnaires. Comparatively, 100% of above average handlers returned at least the first section of the questionnaire, and 90% returned fully completed questionnaires.

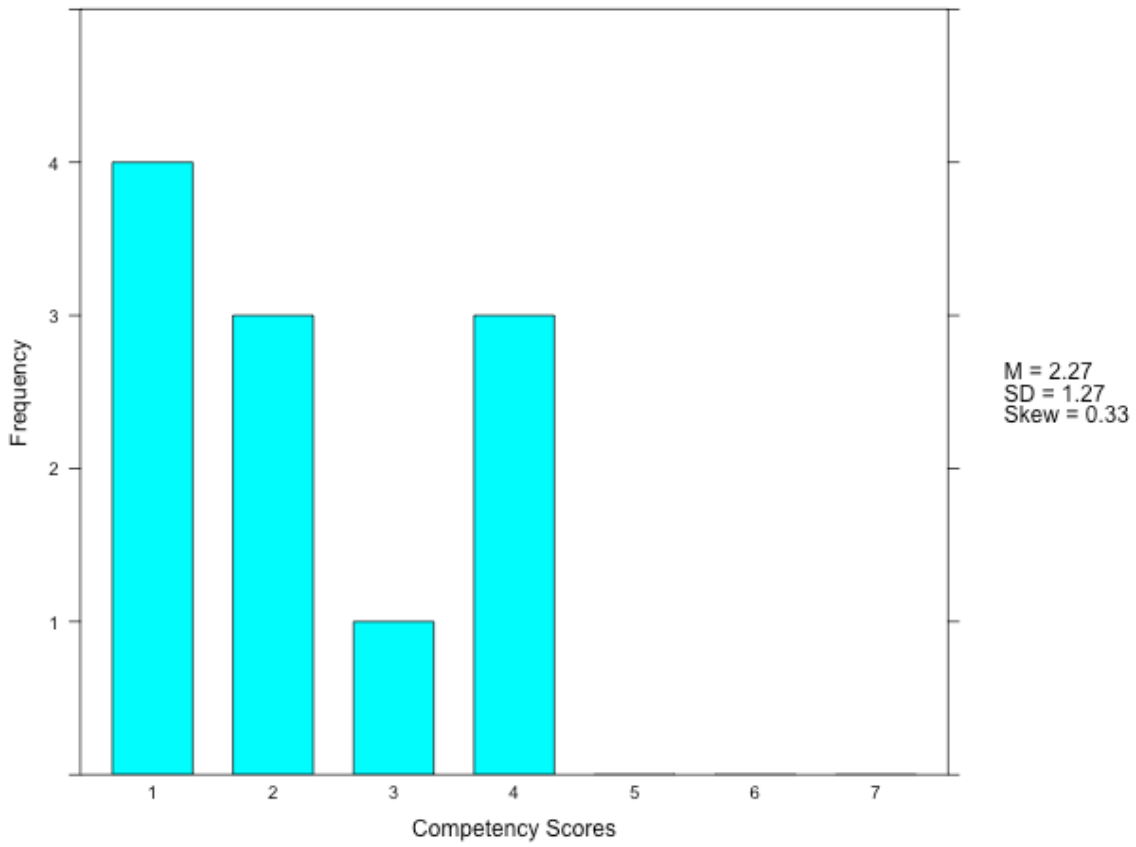


Figure 3.2. Competency score distribution of participants returning no sections of questionnaire ($N = 11$)

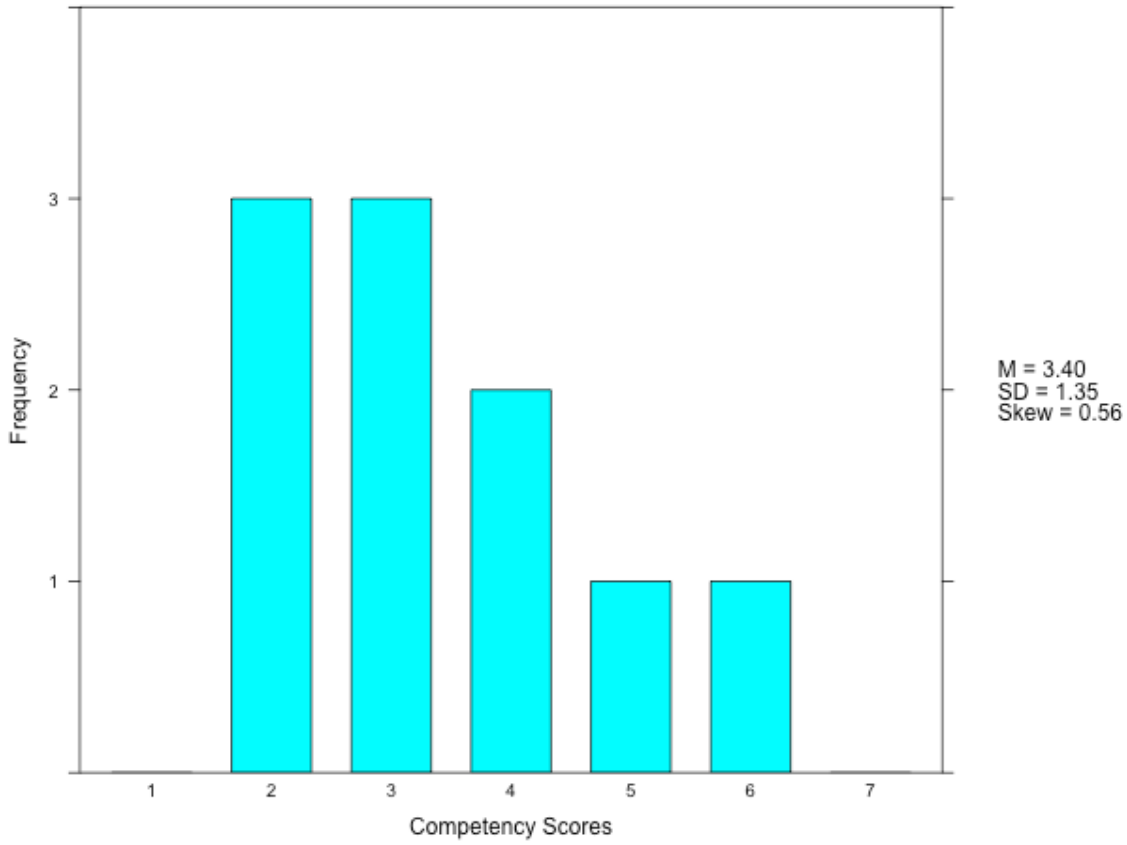


Figure 3.3. Competency score distribution of participants returning only Section 1 of questionnaire ($N = 10$)

Demographics

Forty-six of the Sample59 participants held the position of level 1 handler (78%). Thirteen of the participants held the level 2 team leader position (22%). Figures 3.4 and 3.5 show competency score distributions for each job position. Level 2 team leaders were given overall higher scores (M = 5.38, SD = 1.85) than level 1 handlers (M = 3.52, SD = 1.60; $t_{(64)} = 10.65$, $p < .001$; Figure 3.6.

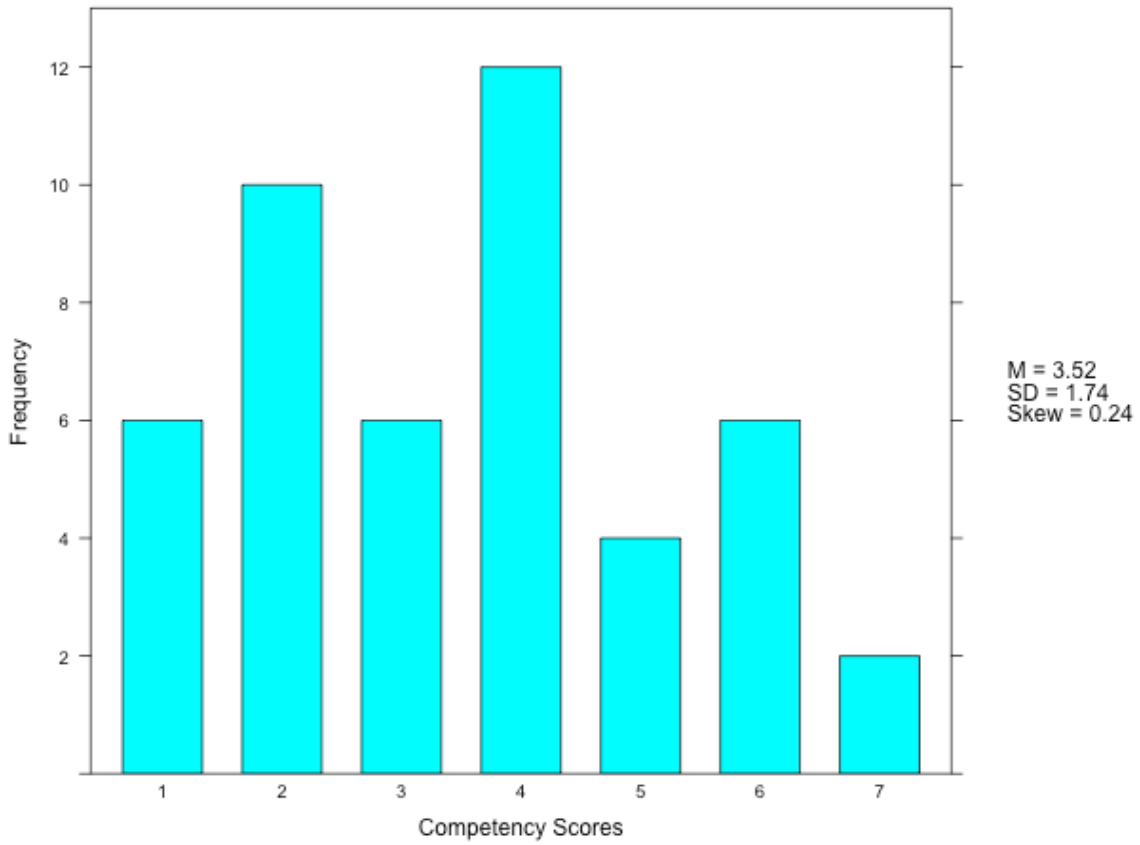


Figure 3.4. Competency score distribution of level 1 handlers (N = 46)

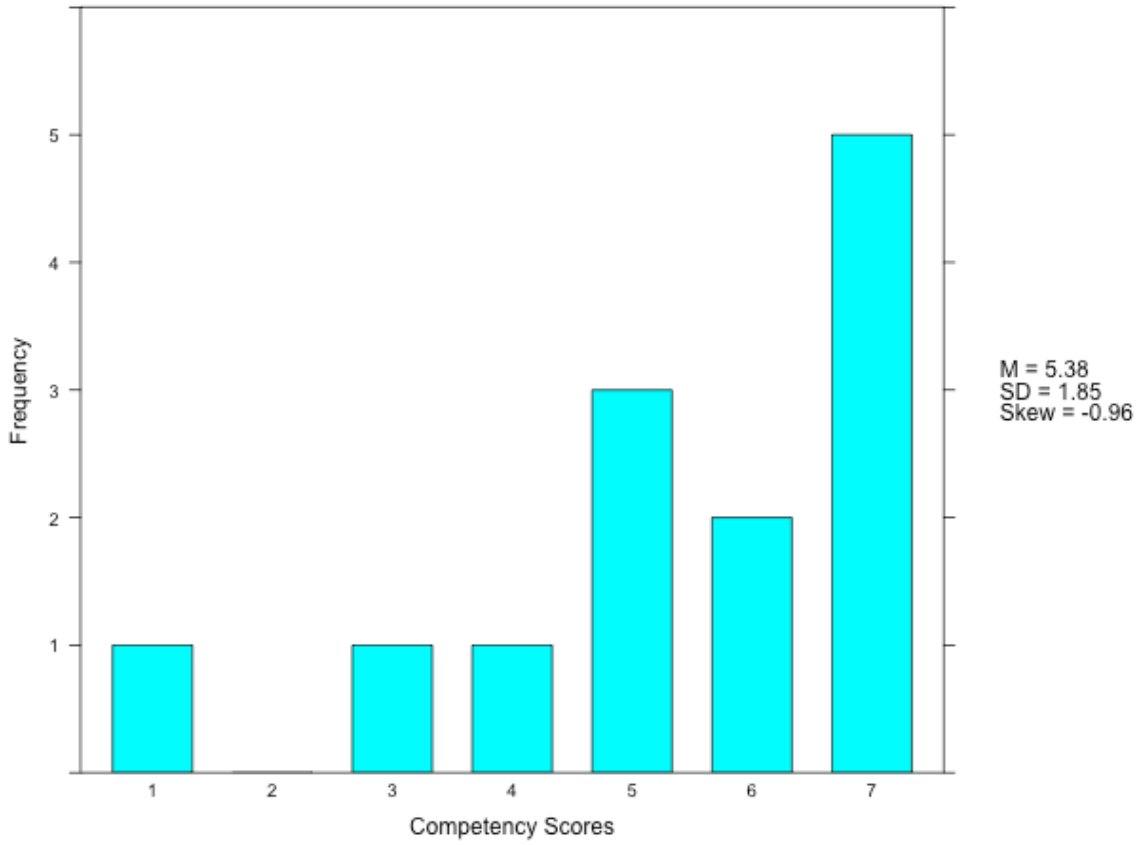


Figure 3.5. Competency score distribution of level 2 team leaders ($N = 13$)

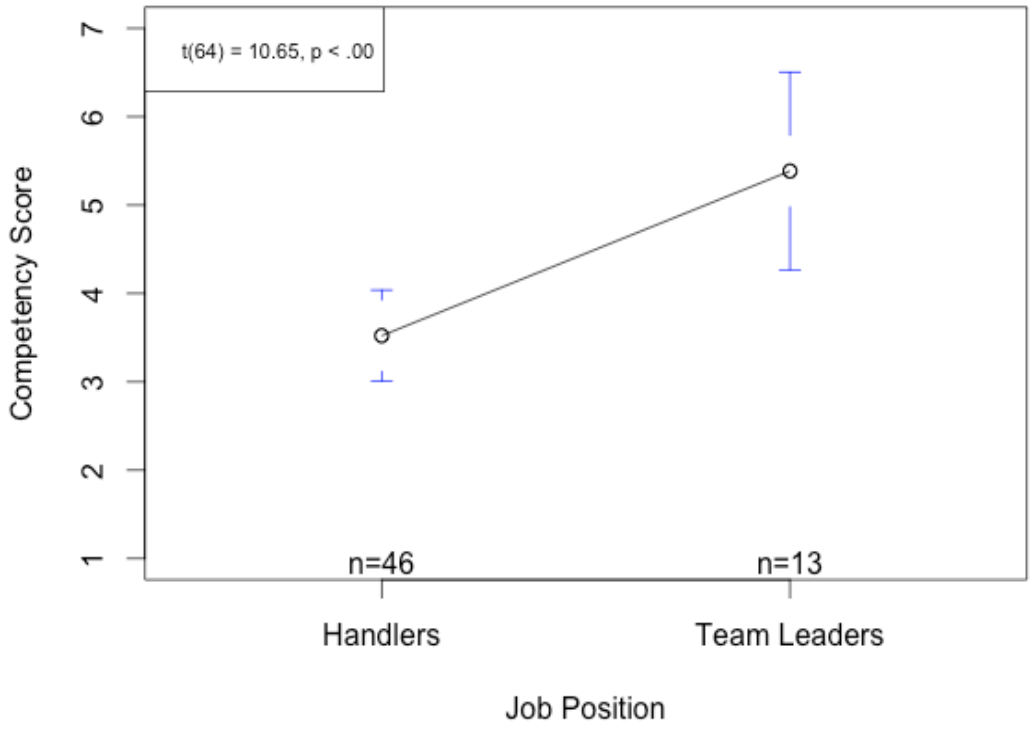


Figure 3.6. Competency score means of level 1 handlers and level 2 team leaders

Sample48

The distribution of Sample48 competency scores (M = 4.31, SD = 1.84) is shown in Figure 3.7. Summaries of results for all linear models fitted to Sample48 are shown in Table 3.5.

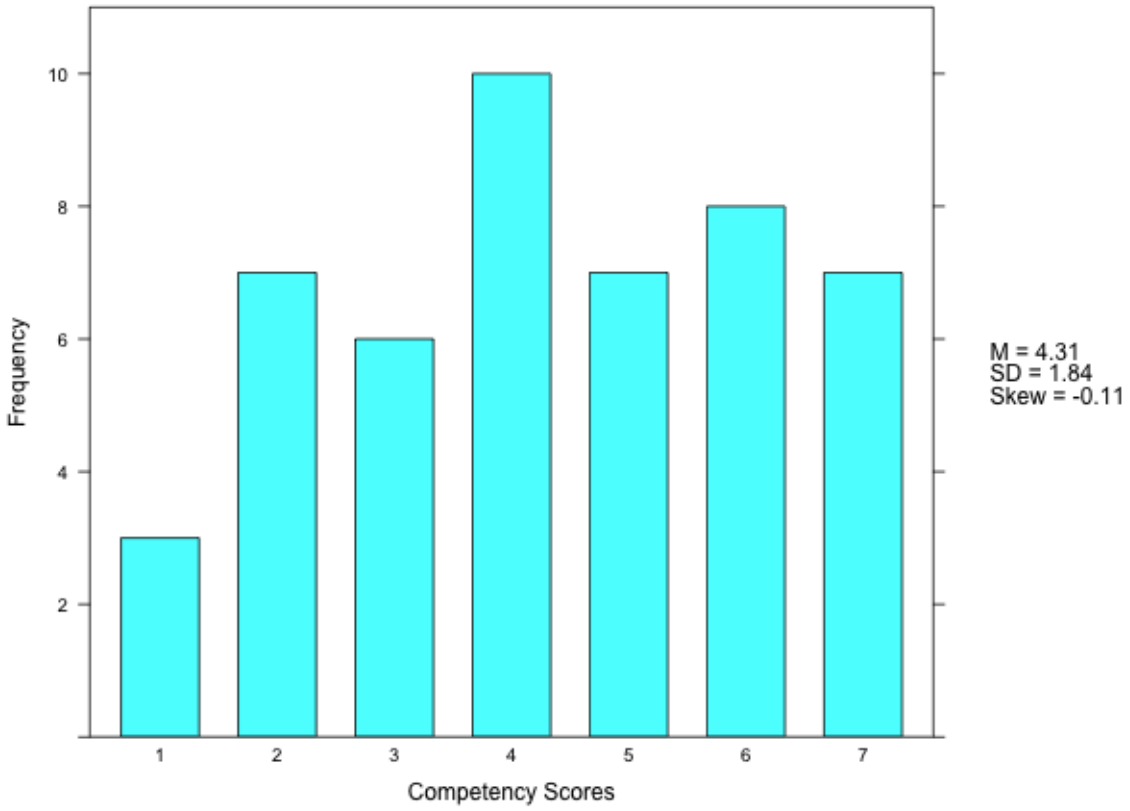


Figure 3.7. Competency score distribution of Sample48 (N = 48)

Table 3.5. Summary of Regression Analysis for CS47 models ($N = 47$) derived from Sample48

	DEM			SPRT			CL-SPRT		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Age	-0.06	0.07	0.15	-0.09	0.07	-0.21	-0.10	0.07	-0.23
Job Position	1.77***	0.59	0.44	1.81***	0.58	0.45	1.88***	0.58	0.47
Gender	0.49	0.67	0.11	0.40	0.65	0.09	0.66	0.65	0.15
Tot Sports				0.11*	0.06	0.24			
CL Sports							0.22**	0.11	0.28
F(df,df)	3.00 (3, 43)			3.07 (4, 42)			3.45 (4, 42)		
R2	0.17			0.23			0.25		
adj R^2	0.12			0.15			0.18		
p	0.04			0.03			0.02		
AICc	190.74			190.27			188.96		
AICc weight	0.20			0.25			0.48		

* p < 0.1, ** p < 0.05, *** p < 0.01

Demographics

Thirty-five level 1 handlers and 13 level 2 team leaders completed either section 1 only or both sections 1 and 2 of the questionnaire. The mean age of Sample48 participants was 34.77 years (SD = 4.13). Thirty-nine were male (81%) and nine were female (19%).

Overall, DEM47 was a significant predictor of competency score ($F_{(3,43)} = 3.00$, p = .04), but failed to explain a large proportion of the variance (R^2 = .17, *adj* R^2 = .12.) DEM47 results showed a positive relationship between job seniority and competency score (b = 1.77, p < 0.01), as we predicted. Our expectation that age would similarly

predict competency scores went unsupported in our analysis (b = -0.06, p = 0.35). The distribution of competency score by age is shown in Figure 3.8.

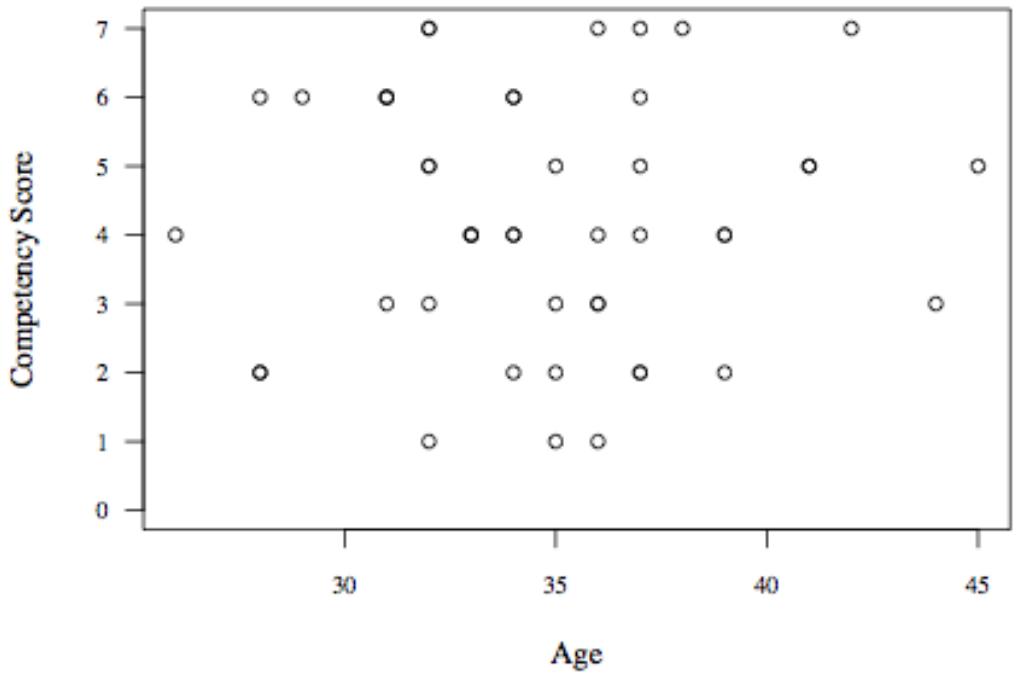


Figure 3.8. Competency score distribution by age (N = 47)

Attitudes towards the dog-human relationship

Of the 48 cases for which we had emotional bond scores, 65% indicated that it was somewhat or extremely important to maintain a strong bond with their detector dogs. Seventeen percent of the participants indicated that it was somewhat or extremely

unimportant to maintain a strong bond. No correlation was found between emotional bond and competency scores (r = - 0.07).

Ninety-five percent of Sample48 indicated that the dog and handler were equally important components to the search team. Due to the lack of variability, no further analysis was done on this item.

Sports Participation

The numbers of total sports played at any level ranged from 0 to 16 (M = 5.48, SD = 4.08). The types of sports played and the distribution of total and competition-level participation is shown in Appendix Q. Appendix R shows sports that were not included in our preselected sport list and were written in by participants selecting the "Other" option.

All participants except one reported having played at least one sporting activity in the past, with 54% of the sample reporting having played two to five different sports. Figure 3.9 shows the distribution of competency scores across the total number of sports each participant played. No one that played over eight total sports received a competency score lower than 4. However, scores ranged across the entire scale for participants playing eight or fewer sports. The number of competition-level sports played ranged from zero to eight (M = 2.56, SD = 2.31). Figure 3.10 shows the distribution of competency scores across the number of competition-level sports played. Seventy-five percent of Sample48 reported participating in at least one competition-level sport. Twenty-one participants (44%) reported playing three or more competition-level sports. Of these 21 participants, only three (15%) received below average competency scores. Twenty-seven participants (57%) reported playing two or fewer competition-level sports. Of these 27 participants, 13 (48%) received below average competency scores.

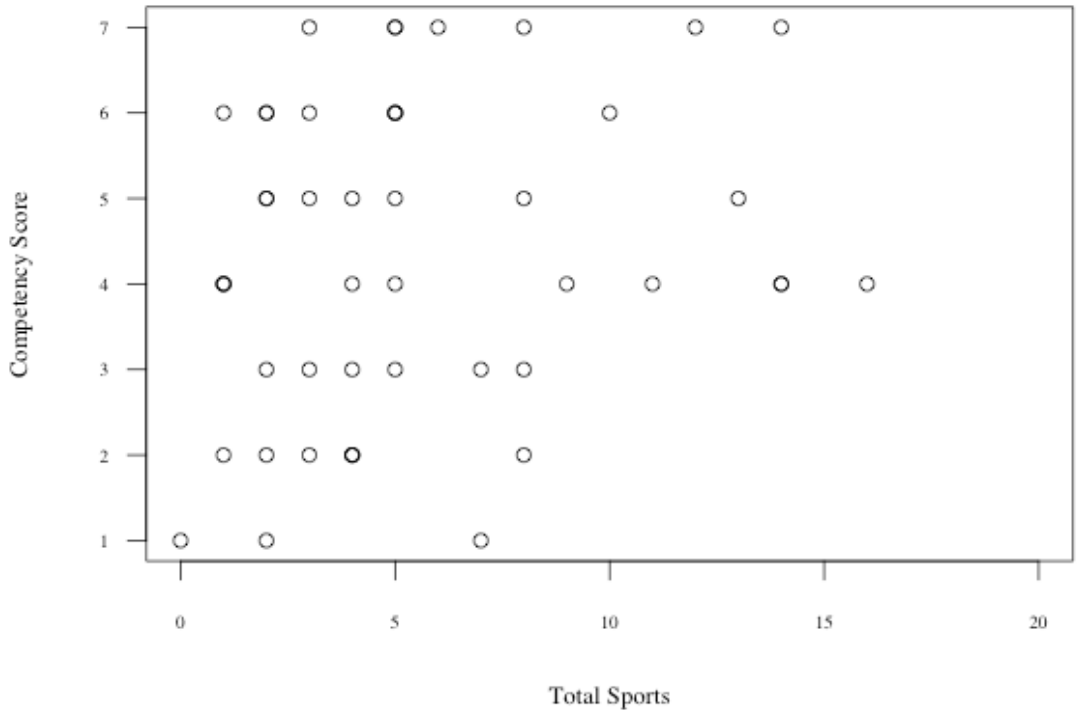


Figure 3.9. Competency score by total sports played at any level ($N = 48$)

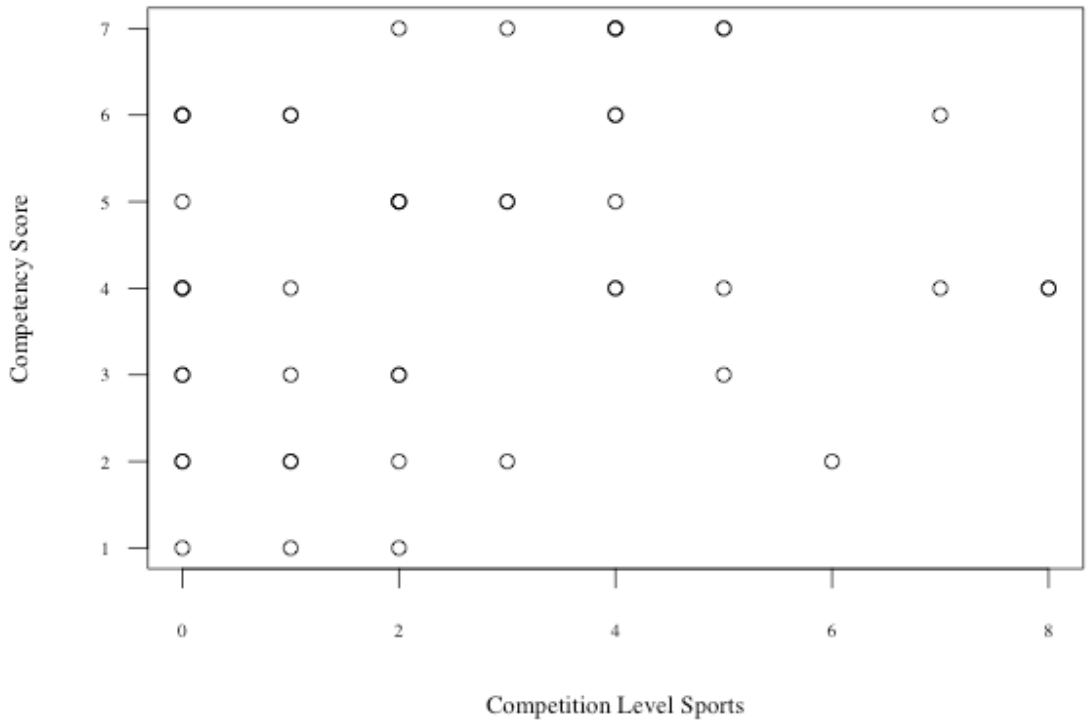


Figure 3.10. Competency score by sports played at competition levels, ($N = 48$)

The ALL-SPRT47 model ($R^2 = 0.23$, adj $R^2 = 0.15$, $F_{(4,42)} = 3.07$, $p < 0.05$) indicated a positive, but non-significant trend of total sports as a predictor of competency ($b = 0.11$, $p = 0.10$). The overall model significance of ALL-SPRT47 was likely driven by the significance of job position ($b = 1.88$, $p < 0.01$). The CL-SPRT47 model ($R^2 = 0.25$, adj $R^2 = 0.18$, $F_{(4,42)} = 3.45$, $p < 0.05$) showed a statistically significant association between competition-level sports and competency score ($b = 0.22, p < 0.05$).

Model comparisons

The CS47 comparison set consisted of DEM47, ALL-SPRT47, CL-SPRT47, and REF47 models. Results (Table 3.4) showed the CL-SPRT47 model (*AICc* = 188.96) to be the strongest model of this set. The total sports variable in ALL-SPRT47 (*AICc* = 190.27) did not result in a meaningful improvement over DEM47 (*AICc* = 190.74). All three models were better predictors than REF47 (*AICc* = 192.49). AICc weights indicated that the .48 relative likelihood of the CL-SPRT47 was higher than the likelihood of 0.25 of the ALL-SPRT47 model and 0.20 of the DEM47 model. REF47 showed a 0.08 probability of being the best model.

When we divide the *AICc* weights, we see that the competition-level sports variable makes the CL-SPRT47 model 2.43 times as likely as DEM47 and 5.83 times as likely as REF47 of being the best model. ALL-SPRT47 is 3.04 times as likely as REF47, but only 1.27 times as likely as DEM47 of being the best model. The DEM47 model is 2.4 times as likely as REF47 of being the best fitting model.

Sample38

The distribution of Sample38 competency scores ($M = 4.55$, $SD = 1.88$) is shown in Figure 3.11. Summaries of results for all linear models fitted to Sample38 are shown in Table 3.6.

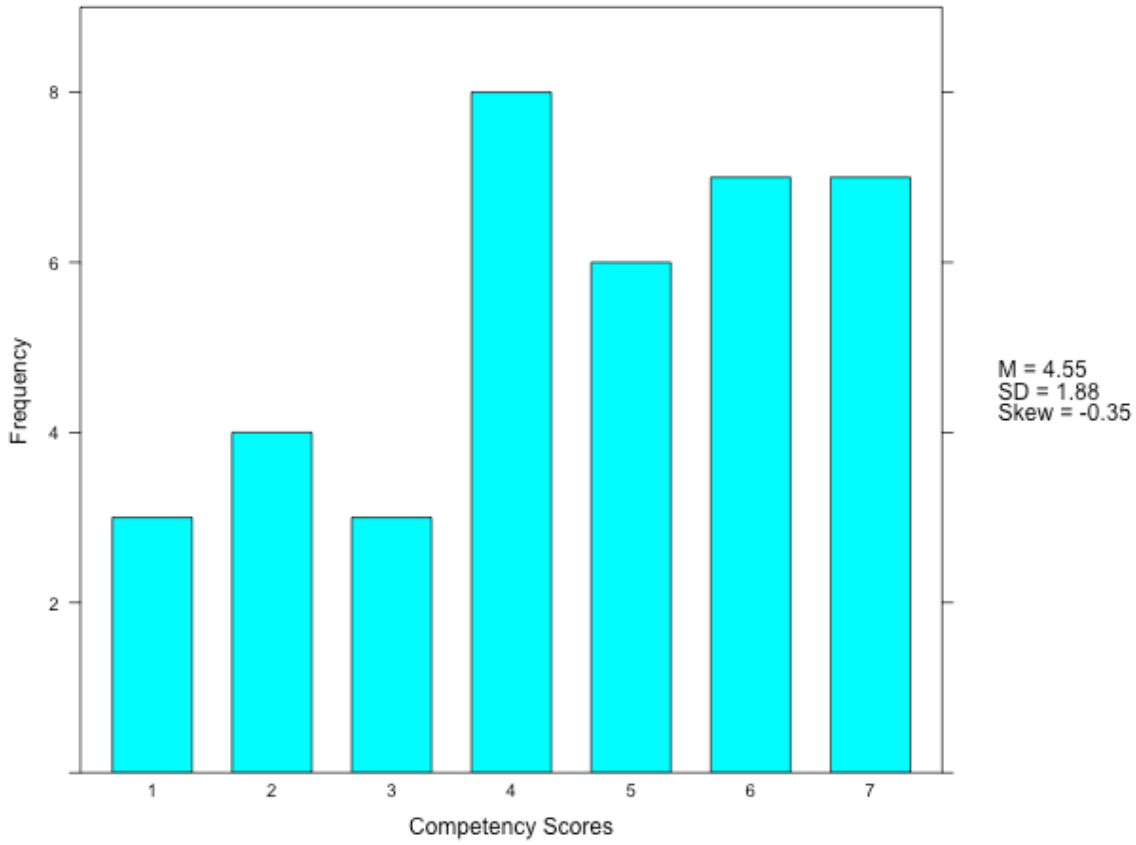


Figure 3.11. Competency score distribution of Sample38 (N = 38)

Table 3.6. Summary of regression analysis for CS37 models ($N = 37$) derived from Sample38

	DEM			PERS			ATT			SPRT			CL-SPRT		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Age	-0.01	0.09	-0.03	-0.01	0.09	-0.01	-0.02	0.11	-0.04	-0.04	0.09	-0.08	-0.04	0.09	-0.08
Job	1.44**	0.67	0.37	1.26*	0.73	0.32	1.55**	0.72	0.40	1.52**	0.67	0.40	1.62**	0.67	0.42
Gender	0.54	0.75	0.12	0.92	0.85	0.21	0.48	0.78	0.11	0 .47	0.75	0.10	0.65	0.74	0.15
IM				-0.15	0.09	-0.34									
O				-1.37*	0.78	-0.29									
C				-0.08	1.01	-0.02									
E				0.20	0.76	-0.05									
A				1.45	0.93	0.31									
N				0.58	0.77	0.18									
Pet Att							0.13	0.40	0 .06						
Bond							-0.29	0.53	-0.10						
Tot. Sprt										0.08	0.08	0.17			
Cl. Sprt													0.19	0.13	0.24
F(df,df)	1.83 (3, 33)			1.74 (9, 27)			1.13 (5, 31)			1.63 (4, 32)			1.98 (4,32)		
R^2	0.14			0.35			0.15			0.17			0.20		
adj R^2	0.06			0.13			0.02			0.07			0.10		
p	0.16			0.16			0.37			0.19			0.12		
AICc	156.36			166.74			161.81			158.06			156.73		
AICc wt.	0.20			0.00			0.01			0.09			0.17		

* p < 0.10, ** p < 0.05, *** p < 0.01

Demographics

Twenty-five level 1 handlers and 13 level 2 team leaders completed both sections of the questionnaire. Thirty of these participants were male (79%) and 8 were female (21%). The mean age of Sample38, excluding the participant who did not reveal his age, was 35.0 ($SD = 3.90$).

Results for the DEM37 model indicated an association between competency and job seniority ($b = 1.44$, $p = 0.04$) with no effect of age ($b = -0.01$, $p = 0.86$), but the overall model was not significant ($R^2 = 0.14$, $adj\ R^2 = 0.06$, $F_{(3,33)} = 1.83$, $p = 0.16$). There were no significant differences ($b = -0.01$, $p = 0.48$) between male and female competency scores (male competency score: $M = 4.29$, $SD = 1.72$; female competency score: $M = 4.11$, $SD = 3.01$).

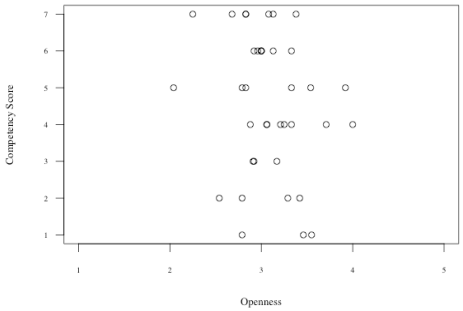
Personality

Conscientiousness scores ranged from 2.83 to 4.88 ($M = 4.04$, $SD = 0.49$). Neuroticism scores ranged from 1.38 to 4.08 ($M = 2.21$, $SD = 0.58$). Openness scores ranged from 2.04 to 4.00 ($M = 3.09$, $SD = 0.40$). Agreeableness scores ranged from 2.83 to 4.58 ($M = 3.91$, $SD = 0.40$). Extraversion scores ranged from 2.10 to 4.50 ($M = 3.56$, $SD = 0.47$). Figure 3.12 shows the relationships of the Big Five dimensions with competency scores.

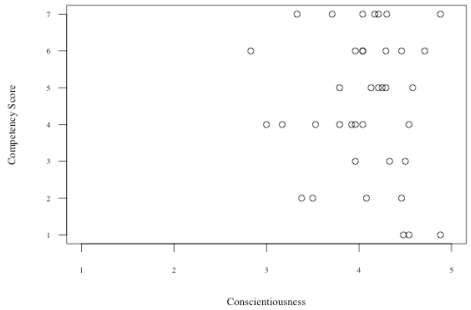
Results of the PERS37 model show that, contrary to our expectations, Conscientiousness ($b = -0.08$, $p = 0.93$) and Neuroticism ($b = 0.58$, $p = 0.46$) were not significantly related to competency score. Openness ($b = -1.37$, $p = 0.09$) and Agreeableness ($b = 1.45$, $p = 0.13$), dimensions that we made no predictions for, showed

stronger correlations than our hypothesized dimensions, but these results did not meet the level of statistical significance. Extraversion (b = 0.20, p = 0.79) was not predictive of competency score. The overall model was not significantly predictive of competency score (R^2 = 0.35, adj R^2 = 0.13, $F_{(9,\ 27)}$ = 1.62, p = 0.16) indicating that overall, an individual's personality was not related to supervisor ratings of competency as a dog handler.

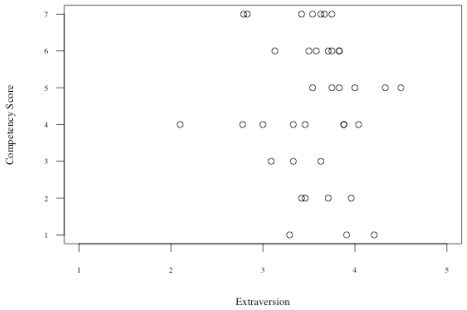
a. Openness



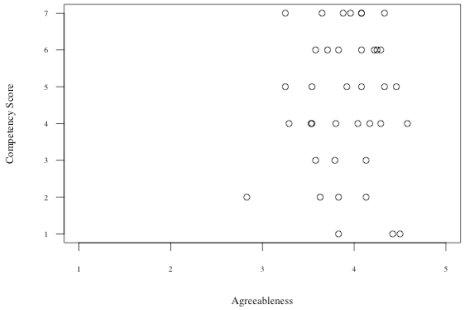
b. Conscientiousness



c. Extraversion



d. Agreeableness



e. Neuroticism

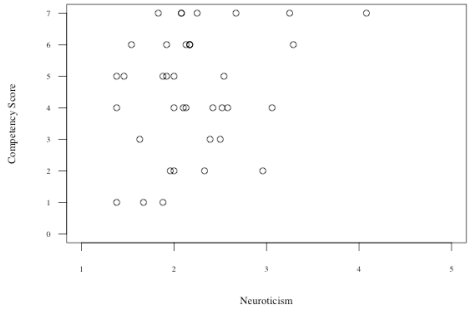


Figure 3.12. Competency score by Big Five personality dimensions, ($N = 38$)

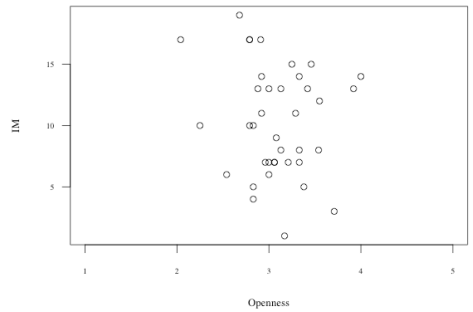
Scatterplots suggested that dishonest answering in the form of impression management may have had an effect on personality results (Figure 3.13). Specifically, a post-hoc linear model regressing the five personality dimensions and DEM variables on IM (Table 3.7) showed that handlers that rated themselves high on Conscientiousness were more likely to use IM ($b = 4.82$, $p < 0.05$). This finding suggests that we may not have been able to make accurate predictions for the Conscientiousness dimension because of inflated Conscientiousness scores as a result of intentional IM by some handlers.

Table 3.7. Regression analysis of Demographics and Personality on Impression Management scores ($N = 37$)

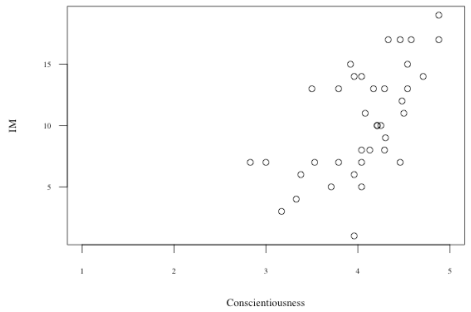
Variable	*B*	*SE B*
Age	-0.04	0.18
Job Position	1.08	1.50
Gender	1.30	1.73
O	-0.29	1.61
C	4.82*	1.87
E	-1.81	1.54
A	2.26	1.89
N	-1.27	1.58
F(df,df)	2.89 (8, 28)	
R^2	0.45	
adj R^2	0.30	
p	0.02	

* $p < 0.05$

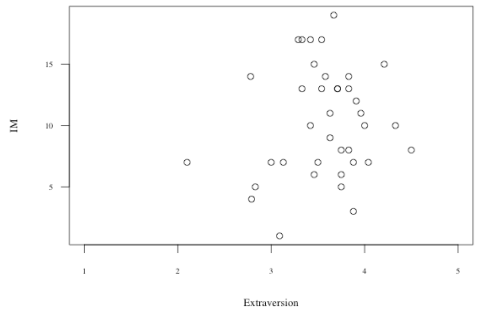
a. Openness



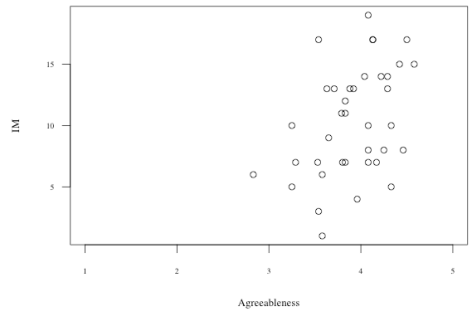
b. Conscientiousness



c. Extraversion



d. Agreeableness



e. Neuroticism

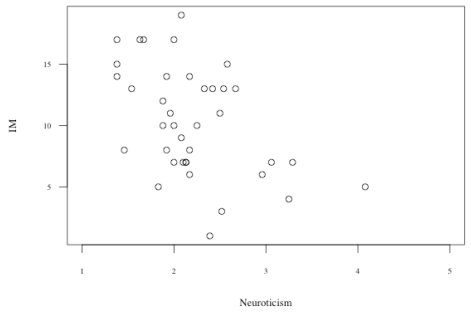


Figure 3.13. Impression management scores by each of the Big Five personality dimensions, (N = 38)

Attitudes toward the dog-human relationship

Attitudes towards pets aggregate scores derived from the PAS-M ($M = 5.20$, $SD = 0.67$) showed low internal reliability ($\alpha = 0.39$) for Sample38. Scores ranged from 3.80 to 6.28. Figure 3.14 shows a lack of variation of pet attitude score compared to competency score.

Fifty-five percent of Sample38 participants agreed that the emotional bond between handler and dog was somewhat important to detection work. An additional 11% believed that the bond was extremely important. Only 13% believed that the bond was either extremely or somewhat unimportant.

ATT37 results show no relationships between attitudes towards the dog-human relationship and competency score ($R^2 = .15$, $adj\ R^2 = .02$, $F_{(5, 31)} = 1.13$, $p = .37$), contrary to our hypothesis. Neither of the predictor variables, pet attitude ($b = -.29$, $p = .59$) and emotional bond ($b = .13$, $p = .75$), were related to competency score.

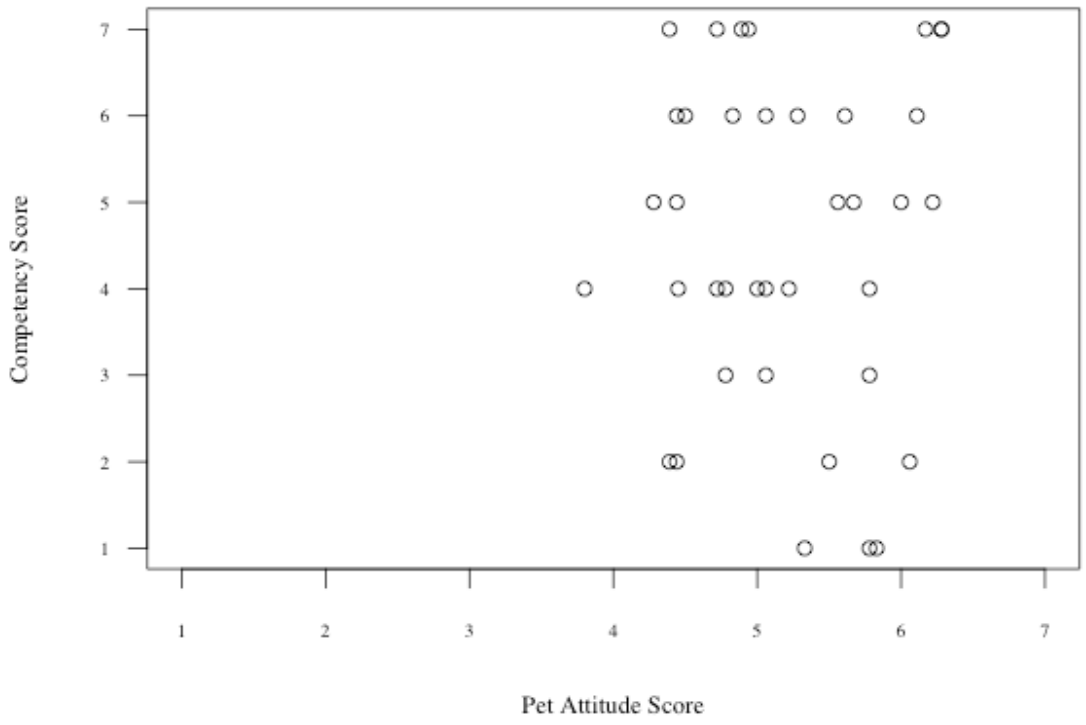


Figure 3.14. Competency score by Pet Attitude score ($N = 38$)

Sports Participation

The total number of sports played at any level ranged from zero to 16 ($M = 5.87$, $SD = 4.09$) for the Sample38 participants. Fifty-five percent reported having participated in between two and five sports.

The ALL-SPRT37 model showed no relationship between total sports play at any level ($b = .08$, $p = .32$) and competency score ($R^2 = .17$, $adj\ R^2 = .07$, $F_{(4, 32)} = 1.63$, $p = .19$).

CL-SPRT37 did not show a strong relationship between competition-level sports ($b = .19, p = .14$) and competency ($R^2 = .20, adj\ R^2 = .10, F_{(4,32)} = 1.98, p = .12$). The total number of sports played at competition-level ranged from zero to eight ($M = 2.92$ ($SD = 2.31$). Fifty percent of participants reported having participated in between two and four sports.

Model Comparisons

CS37 included DEM37, PERS37, ATT37, ALL-SPRT37, CL-SPRT37, and REF37 models (Table 3.5). Our analysis showed DEM37 (*AICc* = 156.36) to be a better predictor than CL-SPRT37 (*AICc* = 156.73), ALL-SPRT37 (*AICc* = 158.06), ATT37 (*AICc* =161.80), and PERS37 (*AICc* = 166.74) models. However, all the models demonstrated worse predictive ability than the REF37 model. The *AICc* weight shows a .52 probability that none of the five predictive models included in CS37 were better fits than the reference model.

When we divide the *AICc* weights, we see that REF37 is 2.58 times as likely as DEM37, the second best model, of being the best fitting model. DEM37 was 1.20 times as likely as CL-SPRT37 and 2.34 times as likely as ALL-SPRT37 meaning that the inclusion of either sports variable reduced the predictive ability of the demographic variables alone. DEM37 was 15.2 times as likely as ATT37, and 179.13 times as likely as PERS37 of being the best fit.

DISCUSSION

Sample59 represents all 59 participants for whom instructors/supervisors had assigned competency scores. We found significant differences in competency score based on the proportion of the questionnaire participants completed. This finding is consistent with the idea that less competent handlers put less effort into the job, and are similarly less likely to put in the effort to complete the questionnaire. Therefore, it may be useful to administer "dummy surveys" or intensive exercises with extended deadlines to job applicants. Results of these surveys or exercises could then be measured for completion and the length of time it takes each participant to complete the task. Applicants completing the tasks closer to the deadline or not fully completing might indicate they are less likely to excel as a detector-dog handler.

The differences in who did and did not complete the full questionnaire may have also impacted the results of our predictive analyses. CS47 models included ten more participants than CS37 models, representing an additional 25% of all below-average handlers. CS47 models that fit demographics and sports participation were all at least twice as likely to be the best fitting models as REF47. CS37 models that fit demographics, personality, attitudes and sports participation, on the other hand, were all inferior to the REF37 model. This pattern suggests that Sample48 and CS47 models include crucial information about below average handlers not present in Sample38 and the CS37 models. Therefore, CS47 models may have been able to more accurately capture differences between high performing and low performing handlers than CS37 models. An alternative explanation is that the competition sports effect is simply not very strong.

Below we consider each of our hypotheses in light of the findings.

Hypothesis 1: Demographics

Results of the DEM47 model found that job seniority was predictive of competency score. Specifically, competency scores of team leaders were expected to be 1.77 higher than entry-level handlers, independent of age and gender variables. The amount of variance explained by DEM47 as a whole was significant and explained 12% of the variance in competency score, when adjusted for the number of predictors. DEM37 showed a similar finding for job seniority, but the full model was not significant and was only able to explain 6% of the variance, when adjusted for the number of predictors. This lack of significance for DEM37 may be due to the lack of information for below average participants when compared to DEM47.

It is not surprising that team leaders received higher competency scores than handlers because it is likely that the best handlers are the ones who get promoted. It is also possible that team leaders were given higher scores due to rater bias, rather than performance. For example, in a study of firefighters and police officers, it was found that seniority alone was an overall poor predictor of performance, and that the commonly used method of promotion based on seniority often had adverse effects (Jacobs, Hofmann, & Kriska, 1990). In other words, assumptions were made about performance based on seniority that did not translate to better future performance. Being older was moderately correlated with being a team leader but we did not find an overall effect of age on competency scores. We predicted an effect of age but our lack of a finding is not surprising based on the inconsistent results of past research on age and job performance (McEvoy & Cascio, 1989).

Hypothesis 2. Personality

Contrary to our expectations, we did not find evidence that Conscientiousness or Neuroticism were associated with competency score in the PERS37 model. Openness and

Agreeableness, two dimensions for which we did not make predictions, showed stronger predictive trends than our hypothesized dimensions, but even these did not reach conventional levels of statistical significance.

Again, because a large portion of the below average participants were not represented in the PERS37, it is possible that personality effects were obscured. In particular, one would expect that individuals low in Conscientiousness would be the most likely participants to return incomplete questionnaires or not complete any questionnaires at all. Unfortunately, we cannot perform this analysis because we only obtained Conscientiousness scores from individuals that completed the entire questionnaire.

An explanation for our lack of a significant personality effect could be that even the least competent handler in our sample was selected over other applicants after undertaking an extensive evaluation and training process. In other words, our sample was subject to considerable restriction of range, consisting of a group of elite detector-dog handlers whose high Conscientiousness and low Neuroticism may have helped distinguish them from other applicants, but not from one another.

Overall, the PERS37 model explained 13% of the variance in competency scores when adjusted for the number of predictors. Six percent of the variance can be attributed to the DEM37 variables so we can conclude that the IM variable and five personality variables explain an additional 7% of the variance. In a post hoc model, we removed the personality variables from the PERS37 model, leaving only the DEM37 variables and the IM variable. The results show that the IM variable alone adds 8% of explained variance to DEM37 (R^2 = .24, *adj* R^2 = .14, $F_{(4, 32)}$ = 1.732, p = .07). This result means that the PERS37 model, after being penalized by *adj* R^2 for the additional number of predictors, explains 1% less of the variance than a model containing only DEM37 and IM. In a post-hoc model removing IM from the PERS37 model, we see that the personality variables

add only an additional 2% to the variance explained by DEM37 (R^2 = 28, *adj* R^2 = .08, $F_{(8, 28)} = 1.39, p = .24$).

Consistent with past studies examining the effects of IM on personality and job performance (Barrick & Mount, 1996; Li & Bagger, 2006) we found IM and Conscientiousness to be positively correlated, but the effects of IM and Conscientiousness on job performance were non-significant. This finding indicates that IM may have artificially inflated the Conscientiousness scores for some participants, and may have obscured any true directional affects of Conscientiousness in our sample.

Hypothesis 3: Attitudes towards the dog-human relationship

We did not see any evidence of a relationship between attitudes towards the dog-human relationship and competency score. When adjusted for the number of predictors, the variance explained by the ATT37 model was only .02. This finding means that when the two attitudes variables are added to the DEM37 model, it causes the model to become worse. There are several possible explanations for our null findings. There was a severe restriction of range with 95% of all participants believing that the roles of the dog and the handler were equally important to an effective detection team. In other words, handlers believe they cannot do their jobs well if the dogs aren't also doing their jobs well. There was also general consensus that the emotional bond between dog and handler was at least somewhat important. Therefore, detector-dog handlers of all competency levels seem to agree on the role of the dogs within odor-detector teams.

The PAS-M was used on the assumption that handlers who held more positive attitudes towards pet dogs would similarly hold more positive attitudes towards their detector dogs. However, this assumption may simply be false. Additionally, the PAS-M's low internal reliability indicates that the scale may not have been an adequate measure for

this population. Due the PAS-M's low internal reliability and lack of even a weak trend in Sample37, it is unlikely that having attitudes measures from the additional below-average participants in Sample47 dataset would have resulted in different PAS-M results.

It is possible that measurements need to be developed that directly assess the relationship quality between detector dogs and professional detector-dog handlers. Past studies have primarily examined companion dogs and their owners as subjects. Typically, pet dog owners personally choose and raise their dogs in their homes as part of the family. Once selected to become active working detector dogs, AC&BPS dogs are moved from the puppy raisers to a central facility and aare generally cared for by AC&BPS kennel staff and not their respective handlers. Such differences may result in different relationship dynamics. Often, relationship quality has been assessed by time spent with the dogs, amount of training, or through surveys aimed for the pet population that may contain items irrelevant to the working dog population. Therefore, simply adapting measures of relationships aimed at pet dogs to working dog contexts may not have been appropriate.

Hypothesis 4: Sports Participation

We found a positive relationship between the total number of sports played and competency score in the ALL-SPRT47 model but the result did not meet the threshold of significance. ALL-SPRT47 explained 15% of the variance in competency score when adjusted for the number of predictors, with 12% being attributable to the DEM47 variables, and 3% to the total sports variable.

The number of sports played at competition-level was found to be a significant predictor in the CL-SPRT47 model. CL-SPRT47 explained 18% of the variance in competency score, with 12% attributable to the DEM47 variables, and 6% to competitive

sports. This finding suggests that correlations between sports involvement and detector-dog handler performance may be moderated by the levels of skill acquired in the sports. It can be reasonably assumed that competition-level participation requires more investment of practice time and repetition than social-level participation, leading to a greater development of the necessary pattern recognition skills involved in each sport. It is possible that, like athletes who can incorporate pattern-recognition skills learned from one sport to another (Baker et al., 2003; Bridge & Toms, 2013), detector-dog handlers are able to incorporate skills acquired from competition-level sports to their jobs. To our knowledge, no research has been conducted into whether pattern-recognition skills learned in sports can later be applied in particular work settings.

The CL-SPRT37 model suggested a similar trend for competition-level sports, but the results were not significant. No effect of total sports was found in ALL-SPRT37.

LIMITATIONS

A number of limitations to the current study should be noted. First, we were unable to obtain more than one competency score for each participant, raising the possibility that the scores were not reliable and valid indicators of competence. Ideally, several raters would independently assign competency scores, which would reduce the likelihood of personal biases affecting the findings. Alternatively, it may be helpful to create objective methods for measuring competency.

A second potential limitation is that the Big Five scores were determined by self-ratings, which can be subject to a number of biases. Supervisor ratings of competency, on the other hand, were distributed based on supervisors past observations of behavior. Past research suggests that observer ratings of Big Five traits may correlate more strongly with supervisor ratings than the self-rating methods used in this study (Hogan, Barrett, & Hogan, 2007).

A third limitation is that the overall effect of competition-level sports experience was not large in Sample47 and was only weakly indicated in Sample37. However, the results are intriguing and should be examined further. More definitive sports measures are recommended. The competition-level and social-level sports variables were rather broad measurements of team and sports expertise, potentially open to varying interpretation among our participants. Past studies have distinguished between expertise levels by contrasting expert professional athletes with experienced but non-expert players (Baker et al., 2003; Bridge & Toms, 2013). It may be beneficial to learn the level of expertise reached by each participant in each sport more precisely. It also may be beneficial to learn whether the types of sports matter. For example, there may be a difference between participation in team sports compared to individual sports. The small sample size made it impossible to meaningfully analyze these types of differences. Future

studies should undertake more detailed examination of the relationships between sports experience, skill acquisition, and dog handler performance.

A fourth potential limitation is that we were able to obtain data from only nine female handlers, and therefore unable to conduct meaningful gender analyses. Collecting data from more female participants would be useful in identifying whether supervisor ratings of job performance may differ based on gender differences.

CONCLUSION

There has been little previous research on the job performance of working-dog handlers. This dearth of studies is despite the fact that working dog-organizations commit a large amount of resources to hiring and training the right personnel. Our study used findings from job-personnel research, companion and working-dog research, sports research, and expert intuition to cast a wide net on potential predictors of job performance in a group of odor-detector-dog handlers. Working dogs and job-personnel studies are currently rich areas of research so it is important to learn how the characteristics of human handlers may contribute to the success or non-success of dog-handler teams. Our study is one of the first to examine this human portion of the dog-human dyad in working contexts.

We recommend the AC&BPS continue to collect demographics, impression management, personality, and sports data from new applicants. Comparisons could then be made between applicants selected for the training program, applicants not selected, and current handlers. It is particularly important for further analysis that more data are collected about below-average handlers. Obtaining more data from female handlers may also be useful to learn about gender effects.

Our strongest finding showed that handlers who completed the entire questionnaire were more likely to be better at their jobs than handlers who did not fully complete the questionnaire. This finding suggests that it may be effective to administer long surveys or online exercises to job applicants and simply measure results for completeness and thoroughness.

A more reliable method for assessing job competency is also recommended. Each handler or applicant should be rated by independent raters and analyzed for reliability. Alternatively, an objective form of job performance assessment should be developed.

Impression management should be looked at further as a potential predictor variable. Our results regarding impression management were not significant but they did indicate a possible trend of dishonest answering being predictive of lower competency.

Openness and Agreeableness dimensions should be analyzed further as potential predictors of competency. Openness showed a negative trend and Agreeableness showed a positive trend for predicting competency, but neither reached statistical significance in our study. However, this result could potentially be strengthened by the inclusion of a larger proportion of below-average handlers.

To increase the efficiency of determining how the quality of the handler-dog relationship may contribute to job performance of odor-detector-dog handlers, it is important that we are able to define the components of a quality relationship. Surveying handlers about their beliefs and what they think makes a quality relationship may be an effective way to develop a clearer definition.

Finally, we recommend that additional sports measures are developed to better assess how skill levels attained in sports might be related to job performance. Asking handlers to score their own sporting skill levels or recording the frequency of play and length of time they participated in each sport may assist in obtaining fine-grained skill-level information. Additionally, we recommend examining whether certain sports may be more effective than others in predicting competency.

Appendix A. Subjective ratings questionnaire items, the sources that influenced item design, and operational definitions used in the current questionnaire form.

Source abbreviation key: HS = Handler Survey, CB – C-BARQ, sCB = shortened C-BARQ, MWD = Military Working Dog Survey, DL = Draft List, DPQ = Dog Personality Questionnaire

Subjective ratings item	Source	Operational Definition
Intelligence/Trainability	HS, CB, sCB	Learns new tasks quickly
Willingness to Please	MWD	Wants to please people
Environmental Soundness	MWD	How well dog copes in different environments to unusual objects, loud noises, and unusual surfaces; confidence
Desire for toy, towel, or ball reward	MWD	Motivated by non-food rewards
Search/Hunt Drive	MWD	Persistence in searching for a toy, ball, or other object that is hidden
Focused/Determined/Task Oriented	MWD	Ability to stay focused when training
Independent	MWD	Ability to work alone without much help from people
General Liveliness/Energy Level/Activity Level	MWD	Dog's general activity level (not only when training)
Hardness	MWD	Lasting effect from an unpleasant or scary experience
Curiosity	MWD	Interest in new or different situations; tendency to approach and sniff the source of stimuli
Excitability	CB, sCB	Tendency to overreact/get overexcited/react in a hyperactive manner
Food motivation	DL	Tendency to try to obtain food
Nosy/uses nose	HS	sniffs objects when out on walks
Stamina	HS	Physical endurance and resistance to fatigue
Playfulness	DPQ	Tendency to display play behavior

Appendix A. *(Continued)*

Kennel Behavior	DL	Tendency to display stress when confined to a kennel
Sociable - towards unknown people	MWD	Tendency to seek interaction with unfamiliar people
Sociable - towards unknown dogs	MWD	Tendency to seek interaction with unfamiliar dogs
Sociable - towards familiar people	MWD	Tendency to seek interaction with familiar people
Sociable - towards unknown dogs	MWD	Tendency to seek interaction with familiar dogs
Human Aggression - familiar	sCB	Tendency to display aggressive behavior toward familiar people
Dog Aggression - familiar	sCB	Tendency to display aggressive behavior toward familiar dogs

Appendix B. Puppy Raiser Subjective Ratings Survey

Your Info:

Name: Sex: Male/Female Age:

Dog Info:

Name: Breed and Color:
Age: Sex: Male/Female Spay/Neuter: Yes/No
How long have you known this dog in months?

How well do you feel you know this dog? 1 – 2 – 3 – 4 – 5 – 6 – 7
Not well at all Very Well

Based on your overall experience of each dog, record your overall impression on the following traits. Make your judgments on the basis of your own understanding of the trait guided by the definitions below. Use your own subjective judgment of typical Detection Dog behavior to decide if the dog you are scoring is above, below, or average for a trait. Use the following scale to make your ratings. Please note that this is not a rating of whether the trait is good or bad, it is a rating of whether the trait is uncharacteristic or characteristic for the dog, in your opinion.

N/A 1----------2----------3----------4----------5----------6----------7	
Trait is extremely **UNcharacteristic** of this dog	Trait is extremely **Characteristic** of this dog

Trait Definition	**Uncharacteristic---- Characteristic**

Appendix B. *(Continued)*

Intelligence/Trainability: *Learns new tasks quickly* • **Low scores:** Slow to learn new tasks, disobedient, stubborn • **High scores:** Quickly learns new tasks, obeys commands	**N/A 1 –2 –3 –4 –5 –6 –7**
Willingness to please: *Wants to please people* • **Low scores**: Not responsive to handler/trainer • **High scores**: Seeks interaction with handler/trainer	**N/A 1 –2 –3 –4 –5 –6 –7**
Environmental Soundness: *How well dog copes in different environments to unusual objects, loud noises, and unusual surfaces; confidence* • **Low scores**: Fearful or scared, easily stressed • **High scores**: Comfortable and confident across environments	**N/A 1 –2 –3 –4 –5 –6 –7**
Desire for toy, towel, or ball reward: *Motivated by non-food rewards* • **Low scores**: Little or no interest in playing with toy/towel/ball; lazy. Does not take toy/dummy/ball in mouth • **High scores**: Will work hard to get toy/towel/ball. May be difficult to get the ball back from the dog	**N/A 1 –2 –3 –4 –5 –6 –7**
Search/Hunt Drive: *Persistence in searching for a toy, ball, or other object that is hidden* • **Low scores**: Loses interest in toy or ball quickly when it is out of sight. Gives up easily when can't find hidden toy or ball. • **High scores**: Highly motivated to search for hidden toy or ball. "Remembers" toy or ball long after it is out of sight and shows strong persistence when hunting for it.	**N/A 1 –2 –3 –4 –5 –6 –7**

Appendix B. *(Continued)*

Focused/Determined/Task Oriented: *Ability to stay focused when training* • **Low scores**: Distractible, non-attentive, gets bored • **High scores**: Focused, undeterred by distractions	**N/A 1 –2 –3 –4 –5 –6 –7**
Independent: *Ability to work alone without much help from people* • **Low scores**: Seeks assistance and reassurance from people • **High scores**: Can work alone, such as will search without assistance from handler, will choose to spend time alone	**N/A 1 –2 –3 –4 –5 –6 –7**
General Liveliness/Energy level/Activity level: *Dog's general activity level (not only when training)* • **Low scores**: Lazy, low energy levels. • **High scores**: Lively, full of energy, constantly moving.	**N/A 1 –2 –3 –4 –5 –6 –7**
Hardness: *Lasting effect from an unpleasant or scary experience* • **Low scores**: Sensitive; takes a long time to get over the experience • **High scores**: Quickly gets over unpleasant or scary situations; does not have a lasting effect	**N/A 1 –2 –3 –4 –5 –6 –7**
Curiosity: *Interest in new or different situations; tendency to approach and sniff the source of stimuli* • **Low scores**: No interest in exploring or investigating novel objects • **High scores**: Extremely likely to investigate novel objects/stimuli	**N/A 1 –2 –3 –4 –5 –6 –7**

Appendix B. *(Continued)*

Excitability: *Tendency to overreact/get overexcited/react in a hyperactive manner* • **Low scores:** Shows very little excitement to sudden or potentially exciting events, like the doorbell ringing, visitors arriving, or seeing new people from inside kennel enclosure • **High scores:** Will bark or yelp hysterically at the slightest disturbance, rushes towards exciting things, gets overexcited during play, difficult to calm down	**N/A 1 –2 –3 –4 –5 –6 –7**
Food motivation: *Tendency to try to obtain food* • **Low scores:** Is not motivated by food, may leave uneaten food in bowl • **High scores:** Persistently begs for food, gulps food quickly	**N/A 1 –2 –3 –4 –5 –6 –7**
Nosy/uses nose: *sniffs objects when out on walks* • **Low scores:** Rarely sniffs at objects when out on walks • **High scores:** Constantly sniffs at objects when out on walks	**N/A 1 –2 –3 –4 –5 –6 –7**
Stamina: *Physical endurance and resistance to fatigue* • **Low scores:** Tires quickly • **High scores:** High physical endurance and resistance to fatigue	**N/A 1 –2 –3 –4 –5 –6 –7**
Playfulness: *Tendency to display play behavior* • **Low scores:** Dog never plays or gets bored quickly during play • **High scores:** Dog enjoys playing with toys, playing tug or fetch	**N/A 1 –2 –3 –4 –5 –6 –7**

Appendix B. *(Continued)*

Kennel Behavior: *Tendency to display stress when confined to a kennel* • **Low Scores:** Dog is relaxed, quiet, often resting in kennel • **High Scores:** Dog often barking, pacing, circling, panting heavily in kennel	**N/A 1 –2 –3 –4 –5 –6 –7**
Sociable – towards unknown people: *Tendency to seek interaction with unfamiliar people* • **Low scores:** Active rejection/retreating • **High scores:** Exaggerated contact seeking/jumping up/licking	**N/A 1 –2 –3 –4 –5 –6 –7**
Sociable -- towards unknown dogs: *Tendency to seek interaction with unfamiliar dogs* • **Low scores**: Active rejection/retreating • **High scores**: Exaggerated contact seeking, playful	**N/A 1 –2 –3 –4 –5 –6 –7**
Sociable – towards familiar people: *Tendency to seek interaction with familiar people* • **Low scores:** Active rejection/retreating • **High scores:** Exaggerated contact seeking/jumping up/licking	**N/A 1 –2 –3 –4 –5 –6 –7**
Sociable -- towards familiar dogs: *Tendency to seek interaction with familiar/family dogs* • **Low scores**: Active rejection/retreating • **High scores**: Exaggerated contact seeking, playful	**N/A 1 –2 –3 –4 –5 –6 –7**
Human aggression – familiar: *Tendency to display aggressive behaviour toward familiar people* • **Low scores:** No visible signs of aggression • **High scores:** Snapping, biting, or attempting to bite when family member verbally corrects, takes away toy or food bowl, approaches the dog while eating, bathing or grooming, stepping over, directly staring at, or at any other time	**N/A 1 –2 –3 –4 –5 –6 –7**

Appendix B. *(Continued)*

Dog aggression – familiar: *Tendency to display aggressive behaviour toward familiar dogs* • **Low scores:** No visible signs of aggression • **High scores:** Snapping, biting, or attempting to bite when approached by a familiar dog while eating, chewing a favourite toy, sleeping, or at any other time	**N/A 1 –2 –3 –4 –5 –6 –7**
Human aggression – unknown people: *Tendency to display aggressive behaviour toward unknown people* • **Low scores:** No visible signs of aggression • **High scores:** Snapping, biting, or attempting to bite when family member verbally corrects, takes away toy or food bowl, approaches the dog while eating, bathing or grooming, stepping over, directly staring at, or at any other time	**N/A 1 –2 –3 –4 –5 –6 –7**
Dog aggression – unknown dogs*:* *Tendency to display aggressive behaviour toward unknown dogs* • **Low scores:** No visible signs of aggression • **High scores:** Snapping, biting, or attempting to bite unknown dogs	**N/A 1 –2 –3 –4 –5 –6 –7**
Prey Drive: *Tendency to chase potential prey* • **Low scores:** Shows no interest in chasing squirrels, cats, bicycles • **High scores:** Would chase squirrels, cats or bicycles if given the opportunity	**N/A 1 –2 –3 –4 –5 –6 –7**
Separation Anxiety/Isolation Distress: *Tendency to show signs of anxiety when left alone or separated from people* • **Low scores:** Relaxed and calm when left alone • **High scores:** When left alone, may excessively bark, whine, scratch at gate, attempt to escape kennel enclosure	**N/A 1 –2 –3 –4 –5 –6 –7**

Appendix B. *(Continued)*

Trait Definition	**Uncharacteristic---- Characteristic**
Do you think this dog will make a good detector dog?	**N/A 1 –2 –3 –4 –5 –6 –7**

List any additional traits you feel should have been included (and define them) and any other comments.

Appendix C. Puppy Raiser Introduction and Study Description

Our department at the University of Texas at Austin is conducting several studies on the behavior and development of working dogs. As part of this research, we are working with Australian Customs and Border Protection Service to learn more about what personality traits are most important to becoming a successful detector dog.

In this packet, you will find four copies of the Puppy Raiser Survey. This survey consists of a series of items designed to capture your personal opinion on specific traits of the dog you are currently raising. We are testing these surveys to eventually use them as a tool to maximize the amount of information we get from the people that know the dogs best!

You will also find an instructional DVD containing important information you need to know before you begin.

Before completing the survey there are a few very important things you need to know:

1. It is very important that *two or more* household members very familiar with the dog's behavior each fill out a separate survey *independently.* We have included four copies of the survey in this packet. If you would like additional copies, please contact Stephen at 043-115-5487 or stephen.debono@mail.utexas.edu.

2. Score each trait using your own subjective opinion, with no help from others.

3. Carefully read the written instructions included on the Puppy Raiser Survey sheet.

4. Answer all items. Circle N/A for 'Not Applicable' if an item does not apply to your situation, or you don't know.

5. Be sure to review and understand the scoring scale.

6. Clearly circle each chosen score.

7. Clearly cross out any score you want to change and circle the new score.

8. Only circle one score per item.

9. Please review the short instructional video on the included DVD before beginning the survey.

If you have any questions or would like to discuss any aspect of this study, please do not hesitate to contact Stephen DeBono at 043-115-5487 or through e-mail at stephen.debono@mail.utexas.edu.

The University of Texas at Austin and the AC & BPS greatly appreciate you taking the time to complete the Puppy Raiser Survey. Your assistance goes a long way in helping us learn how to do the best job we can to raise the puppies for success.

Appendix D. Puppy Raiser Free Text Responses

Dog	Rater Comments
Billie	Billie is a brood bitch and a delight to share our family and home with
Coora (Rater 1)	She became the perfect mother, teaches other pups very well, energetic/intelligent/lovable dog!
(Rater 2)	Dog's response to being chastised verbally or the withdrawal of human attention (i.e. to correct by loss of eye contact)
Dory	Quick to learn, eager to please and take part in activities
Fergus	Destroy - high, chew - high
Gali	Not happy around water - hose bath
Igor	Very quiet and placid, would make a great assistance dog. Lovely temperament. Very loving and caring dog.
Kip	With proper training, will make a better than average detector dog
Kita	Loyalty (to puppy raiser/handler)
Kurt	Dog constantly needs attention
Lawson	Dog agg familiar: "play fights with Plato (also lives here) however doesn't seem too aggressive and not while eating etc."
Lester (Rater 1)	Wants to be the dominant dog of all three in the household
(Rater 2)	Loves a cuddle
Meg	Response to command/willingness to do new stuff

Appendix D. *(Continued)*

Oreo (Rater 1)	We don't train the dog to hunt is done by the handler when they come
(Rater 2)	Specialist training done by the dept
Otis	Destruct - destroy high - 7, chewing - 7
Perkin	I could not answer questions about interactions with familiar dogs as we don't see him with other dogs
Plato	Dog agg familiar: "play fights with Lawson (also lives here) however doesn't seem too aggressive and not while eating etc"
Quilla	Persistence/persisting with tasks and level of compliance when asked to try again and again
Quondo	N/A kennel - I don't observe Quondo in the customs environment so cannot comment on his behavior there
Rocco (Rater 1)	intelligence level - Rocco shows a very high level of intelligence
(Rater 2)	very good with children and adults
Rolf	Transportability? Eagerness to travel in car
Sarge	I feel it's a bit early to determine if he will make a good detector dog
Victor (Rater 1)	just starting to show positive traits, carrying and tug of war
(Rater 2)	loves to chase birds - groans when he is picked up via front legs, unsure if in pain or just grumpy - maturity is showing improvements in his carrying traits - he is very lazy when he comes in. Happy to lay on the couch
Wade	attention seeking - wants peoples attention on him, to play

Appendix D. *(Continued)*

Watson	Watson was a slow starter when first arriving. In recent weeks he continues to surprise with his hidden capabilities. He is progressing well, particularly outside the yard. Does not take well ti the water at beach. Happy to share sleeping space with York :)
Webby	This dog has been a detector dog
Wilma	loves to be close to people. Will often be happy sitting on the mat with people in the same room. Very placid
Xavier (Rater 1)	tries to be center of attention, a bit pushy but not in an aggressive way
(Rater 2)	he was very slow in the beginning but came out of his shell in a big way - very eager to please
Yogi	mischievous - must be kept busy. Mischievous when left alone. This dog in my opinion is very determined and must have things go his way
Yona	ability to jump.
York (Rater 1)	York has been a pleasure to care for from day 1. Though he will be missed when he returns to customs full time. I wish both York and his future handler much success. (Very tolerant of Watson)
(Rater 2)	York seeks contact but more to play with him not pet. Very down to business. Calm but very motivated to play. I know this is our first dog but we have felt he is an exceptional dog. Hopefully he will go far

Appendix E. Predicted outcomes and classifications of all subjects using Leave One Out Cross Validation (0.50 threshold)

Subject	Predicted probability of successful outcome	Predicted label	True outcome
Bailey	0.78	Pass	Pass
Banjo	0.82	Pass	Pass
Billie	0.85	Pass	Pass
Cooper	0.73	Pass	Fail
Coora	0.84	Pass	Pass
Dobby	0.70	Pass	Fail
Dolly	0.74	Pass	Pass
Dory	0.83	Pass	Pass
Elise	0.68	Pass	Fail
Elmo	0.70	Pass	Pass
Elsa	0.79	Pass	Fail
Farrah	0.83	Pass	Pass
Fergus	0.81	Pass	Fail
Fern	0.82	Pass	Pass
Flicka	0.73	Pass	Pass
Floyd	0.74	Pass	Pass
Gali	0.82	Pass	Pass
Gilly	0.74	Pass	Pass
Griff	0.74	Pass	Pass
Gunda	0.70	Pass	Pass
Halle	0.75	Pass	Fail
Hawk	0.72	Pass	Pass
Honey	0.82	Pass	Pass
Igor	0.65	Pass	Fail
Ike	0.87	Pass	Pass
Ilka	0.76	Pass	Pass
Indi	0.66	Pass	Pass
Ivory	0.69	Pass	Pass
Jet	0.79	Pass	Pass
Keogh	0.78	Pass	Pass
Kiesha	0.84	Pass	Pass
Kimba	0.79	Pass	Pass
Kingy	0.65	Pass	Pass
Kip	0.58	Pass	Pass

Appendix E. *(Continued)*

Kita	0.75	Pass	Pass
Koda	0.69	Pass	Pass
Kurt	0.56	Pass	Fail
Lawson	0.72	Pass	Fail
Lester	0.69	Pass	Pass
Libby	0.73	Pass	Fail
Maddie	0.82	Pass	Pass
Marty	0.72	Pass	Pass
Maya	0.80	Pass	Pass
Meg	0.65	Pass	Pass
Molly	0.81	Pass	Pass
Onnie	0.74	Pass	Pass
Onya	0.74	Pass	Fail
Oreo	0.66	Pass	Fail
Oscar	0.58	Pass	Pass
Otis	0.84	Pass	Pass
Pablo	0.84	Pass	Pass
Perkin	0.78	Pass	Pass
Plato	0.81	Pass	Pass
Polly	0.72	Pass	Pass
Quebec	0.79	Pass	Pass
Queenie	0.68	Pass	Fail
Quiana	0.74	Pass	Pass
Quilla	0.80	Pass	Pass
Quondo	0.75	Pass	Fail
Rocco	0.76	Pass	Pass
Rolf	0.76	Pass	Pass
Sarge	0.77	Pass	Fail
Spike	0.74	Pass	Fail
Timmy	0.66	Pass	Pass
Trinny	0.81	Pass	Pass
Ultra	0.75	Pass	Fail
Una	0.75	Pass	Fail
Uno	0.47	Fail	Pass
Unwin	0.57	Pass	Pass
Urban	0.72	Pass	Fail
Uriah	0.73	Pass	Pass

Vega	0.43	Fail	Pass

Appendix E. *(Continued)*

Victor	0.25	Fail	Fail
Villian	0.71	Pass	Pass
Virgil	0.37	Fail	Pass
Vixen	0.84	Pass	Pass
Wade	0.84	Pass	Pass
Wally	0.73	Pass	Pass
Watson	0.40	Fail	Pass
Wattle	0.76	Pass	Pass
Webby	0.82	Pass	Pass
Wilbur	0.37	Fail	Pass
Wilma	0.27	Fail	Fail
Wilson	0.61	Pass	Pass
Winnie	0.67	Pass	Fail
Wolf	0.73	Pass	Pass
Wynn	0.75	Pass	Fail
Xanthos	0.81	Pass	Pass
Xavier	0.74	Pass	Fail
Xia	0.67	Pass	Fail
Xing	0.79	Pass	Fail
Xion	0.65	Pass	Pass
Yasmine	0.79	Pass	Fail
Yella	0.77	Pass	Pass
Yogi	0.66	Pass	Pass
Yona	0.81	Pass	Pass
Yoomee	0.79	Pass	Pass
York	0.84	Pass	Pass
Zeb	0.80	Pass	Pass
Zeplin	0.74	Pass	Pass
Zeta	0.65	Pass	Fail
Zinta	0.74	Pass	Pass
Zoltan	0.74	Pass	Pass
Zorro	0.75	Pass	Fail

Appendix F. IPIP-NEO Big Five Personality Inventory with trait dimension identified for each item

Dimension	Item #	Item Content
Neuroticism	1	Worry about things.
Extraversion	2	Make friends easily.
Openness	3	Have a vivid imagination.
Agreeableness	4	Trust others.
Concientiousness	5	Complete tasks successfully.
Neuroticism	6	Get angry easily.
E	7	Love large parties.
O	8	Believe in the importance of art.
A	9	Use others for my own ends.
C	10	Like to tidy up.
N	11	Often feel blue.
E	12	Take charge.
O	13	Experience my emotions intensely
A	14	Love to help others
C	15	Keep my promises.
N	16	Find it difficult to approach others.
E	17	Am always busy.
O	18	Prefer variety to routine.
A	19	Love a good fight.
C	20	Work hard.
N	21	Go on binges.

Appendix F. *(Continued)*

E	22	Love excitement.
O	23	Love to read challenging material.
A	24	Believe that I am better than others.
C	25	Am always prepared.
N	26	Panic easily.
E	27	Radiate joy.
O	28	Tend to vote for candidates on the political left.[1]
A	29	Sympathize with the homeless.
C	30	Jump into things without thinking.
N	31	Fear for the worst.
E	32	Feel comfortable around people.
O	33	Enjoy wild flights of fantasy.
A	34	Believe that others have good intentions.
C	35	Excel in what I do.
N	36	Get irritated easily.
E	37	Talk to a lot of different people at parties.
O	38	See beauty in things that others might not notice.
A	39	Cheat to get ahead.
C	40	Often forget to put things back in their proper place.
N	41	Dislike myself.
E	42	Try to lead others.
O	43	Feel others' emotions.
A	44	Am concerned about others.
C	45	Tell the truth.
N	46	Am afraid to draw attention to myself.
E	47	Am always on the go.
O	48	Prefer to stick with things that I know.
A	49	Yell at people.

[1] Changed from “Tend to vote for liberal political candidates”

[2] changed from “Tend to vote for conservative political candidates”

Appendix F. *(Continued)*

C	50	Do more than what's expected of me.
N	51	Rarely overindulge.
E	52	Seek adventure.
O	53	Avoid philosophical discussions.
A	54	Think highly of myself.
C	55	Carry out my plans.
N	56	Become overwhelmed by events.
E	57	Have a lot of fun.
O	58	Believe that there is no absolute right or wrong.
A	59	Feel sympathy for those who are worse off than myself.
C	60	Make rash decisions.
N	61	Am afraid of many things
E	62	Avoid contacts with others.
O	63	Love to daydream.
A	64	Trust what people say.
C	65	Handle tasks smoothly.
N	66	Lose my temper.
E	67	Prefer to be alone.
O	68	Do not like poetry.
A	69	Take advantage of others.
C	70	Leave a mess in my room.
N	71	Am often down in the dumps.
E	72	Take control of things.
O	73	Rarely notice my emotional reactions.
A	74	Am indifferent to the feelings of others.
C	75	Break rules.
N	76	Only feel comfortable with friends.
E	77	Do a lot in my spare time.
O	78	Dislike changes.

Appendix F. *(Continued)*

A	79	Insult people.
C	80	Do just enough work to get by.
N	81	Easily resist temptations.
E	82	Enjoy being reckless.
O	83	Have difficulty understanding abstract ideas.
A	84	Have a high opinion of myself.
C	85	Waste my time.
N	86	Feel that I'm unable to deal with things.
E	87	Love life.
O	88	Tend to vote for candidates on the political right. [2]
A	89	Am not interested in other people's problems.
C	90	Rush into things.
N	91	Get stressed out easily.
E	92	Keep others at a distance.
O	93	Like to get lost in thought.
A	94	Distrust people.
C	95	Know how to get things done.
N	96	Am not easily annoyed.
E	97	Avoid crowds.
O	98	Do not enjoy going to art museums.
A	99	Obstruct others' plans.
C	100	Leave my belongings around.
N	101	Feel comfortable with myself.
E	102	Wait for others to lead the way.
O	103	Don't understand people who get emotional.
A	104	Take no time for others.
C	105	Break my promises.

[2] changed from “Tend to vote for conservative political candidates”

Appendix F. *(Continued)*

N	106	Am not bothered by difficult social situations.
E	107	Like to take it easy.
O	108	Am attached to conventional ways.
A	109	Get back at others.
C	110	Put little time and effort into my work.
N	111	Am able to control my cravings.
E	112	Act wild and crazy.
O	113	Am not interested in theoretical discussions.
A	114	Boast about my virtues.
C	115	Have difficulty starting tasks.
N	116	Remain calm under pressure.
E	117	Look at the bright side of life.
O	118	Believe that we should be tough on crime.
A	119	Try not to think about the needy.
C	120	Act without thinking.

Appendix G. Balanced Inventory of Desired Responding Impression Management Scale Vers. 6

Using the scale below as a guide, write a number beside each statement to indicate how true it is.

1	2	3	4	5	6	7
not true			somewhat			very true

1. I sometimes tell lies if I have to.
2. I never cover up my mistakes.
3. There have been occasions when I have taken advantage of someone.
4. I never swear.
5. I sometimes try to get even rather than forgive and forget.
6. I always obey laws, even if I'm unlikely to get caught.
7. I have said something bad about a friend behind his/her back.
8. When I hear people talking privately, I avoid listening.
9. I have received too much change from a salesperson without telling him or her.
10. I always declare everything at customs.
11. When I was young I sometimes stole things.
12. I have never dropped litter on the street.
13. I sometimes drive faster than the speed limit.
14. I never read sexy books or magazines.
15. I have done things that I don't tell other people about.
16. I never take things that don't belong to me.
17. I have taken sick-leave from work or school even though I wasn't really sick.
18. I have never damaged a library book or store merchandise without reporting it.
19. I have some pretty awful habits.
20. I don't gossip about other people's business.

Appendix H. PAS-M – Pet Attitude Scale Modified

Instructions: Please answer each of the following questions as honestly as you can, in terms of how you feel right now. There aren't any right or wrong answers. All that matters is that you express your true thoughts on the subject.

Please answer by selecting of the following seven numbers for each question.

1 Strongly Disagree	2 Moderately Disagree	3 Slightly Disagree	4 Unsure	5 Slightly Agree	6 Moderately Agree	7 Strongly Agree

1. I really like seeing pets enjoy their food.
2. My pet means more to me than any of my friends (or would if I had one).
3. I would like to have a pet in my home.
4. Having pets is a waste of money.
5. House pets add happiness to my life (or would if I had one).
6. I feel that pets should always be kept outside.
7. I spend time every day playing with my pet (or would if I had one).
8. I have occasionally communicated with my pet and understood what it was trying to express (or would if I had one).
9. The world would be a better place if people would stop spending so much time caring for their pets and started caring more for other human beings instead.
10. I like to feed animals out of my hand.
11. I love pets.
12. Animals belong in the wild or in zoos, but not in a home.
13. If you keep pets in the house you can expect a lot of damage to furniture.
14. I like house pets.
15. Pets are fun but it's not worth the trouble of owning one.
16. I frequently talk to my pets (or would if I had one).
17. I hate animals.
18. You should treat your house pets with as much respect as you would a human member of your family.

Appendix I. Attitudes Items

Emotional Bond (EB)

To what extent do you agree with the following statement: It is important for a dog handler to maintain a strong emotional bond with his/her detector dog.

□ 1. agree

□ 2. disagree

□ 3. neither agree nor disagree

□ 4. agree

□ 5. strongly agree

Relative Importance (RI)

Please select the statement you agree with the most.

□ 1. The quality of the dog is the only important part of successful detection work

□ 2. The quality of the dog is the most important part of successful detection work

□ 3. The quality of the dog and the quality of the handler are equally important parts of successful detection work

□ 4. The quality of the handler is the most important part of successful detection work

□ 5. The quality of the handler is the only important part of successful detection work

Appendix J. Sports Items

Do you have experience playing the following sports? (Check all that apply)

Cricket	□	Golf	□
Soccer	□	Rugby	□
Athletics	□	Tennis	□
Basketball	□	Cycling	□
Netball	□	Baseball	□
Volleyball	□	Hockey	□
AFL	□	Horse Riding	□
Squash	□	Gymnastics	□
Swimming	□	Dancing	□
Badminton	□	Other	□
Lacrosse	□		

If "Other" was selected, the following item was presented along with a text box:

Please describe "Other". Be as specific as possible.

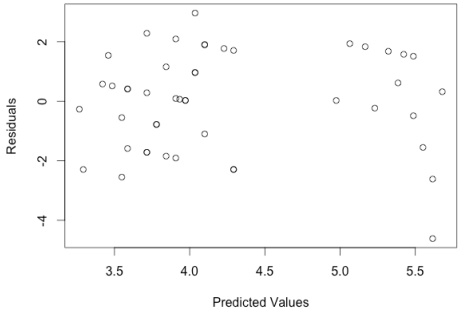
For each sport that was selected, an item in the following format was presented:

Did you play [Sport Name] at a competitive or social level?

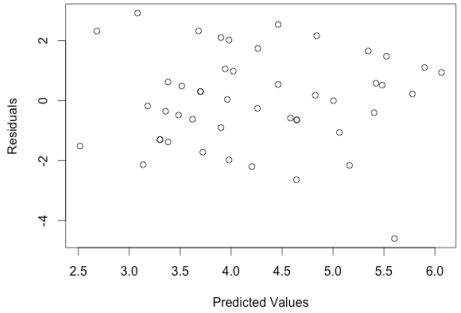
Competitive	□
Social	□

Appendix K. Heteroscedasticity assumption check for CS47 models. Heteroscedasticity assumption is met when the pattern of the plotted residuals is not constant at different levels of the predicted values.

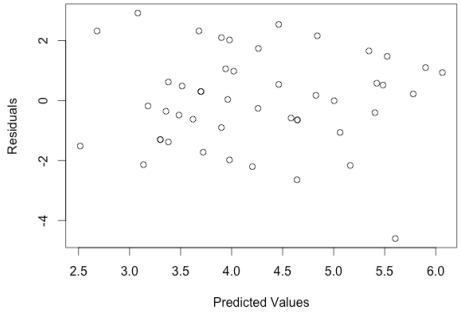
a. DEM47



b. ALL-SPRT47

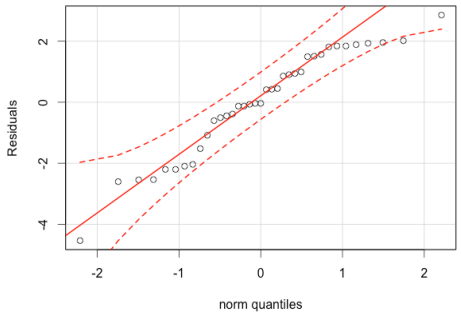


c. CL-SPRT47

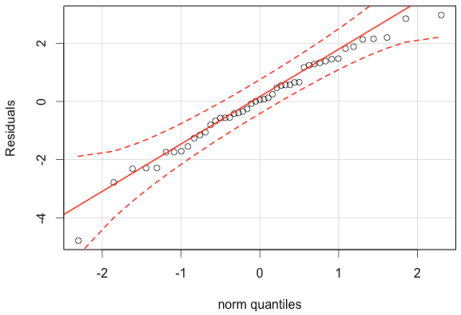


Appendix L. Normality assumption check for CS47 models. QQ plots show model residuals plotted along solid diagonal red line. Normality is violated when excessive numbers of residuals 95% confidence intervals, identified by the red diagonal dotted lines.

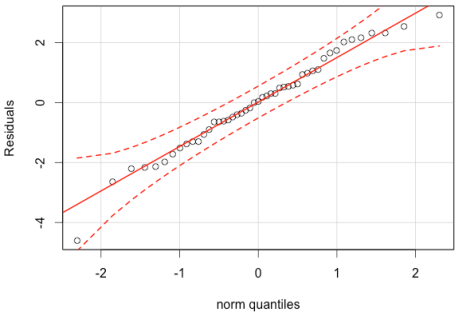
a. DEM47



b. ALL-SPRT47

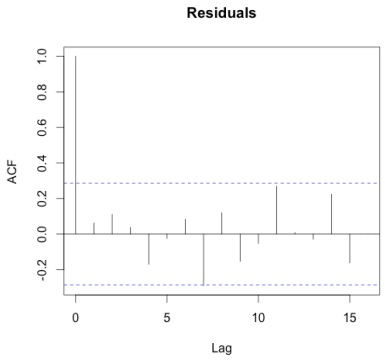


c. CL-SPRT47

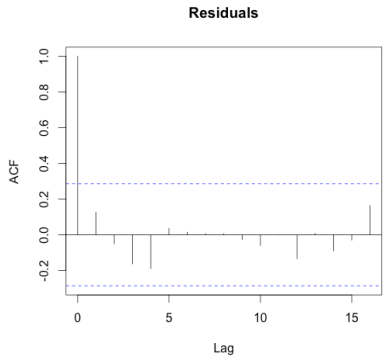


Appendix M. Autocorrelation estimates of CS47 models. Autocorrelation assumptions are met when lags over 0, plotted along x-axis, do not exceed height of dotted horizontal line.

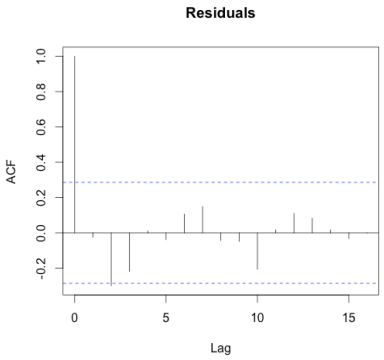
a. DEM47



b. ALL-SPRT47

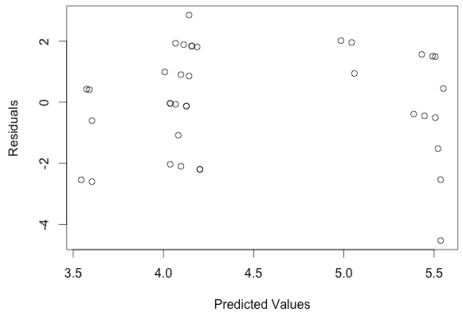


c. CL-SPRT47

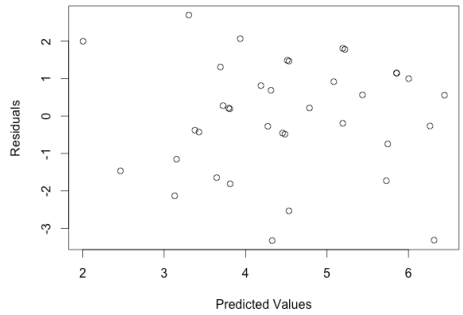


Appendix N. Heteroscedasticity assumption check for CS37 models. Heteroscedasticity assumption is met when the pattern of the plotted residuals is not constant at different levels of the predicted values.

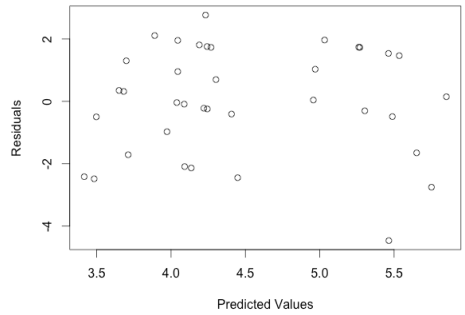
a. DEM37



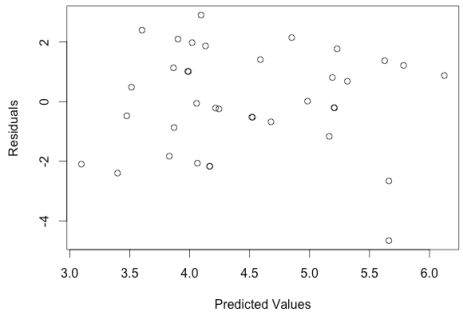
b. PERS37



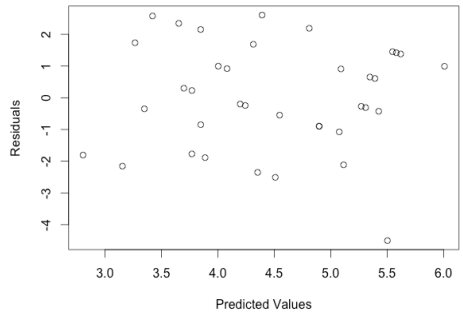
c. ATT37



d. ALL-SPRT37

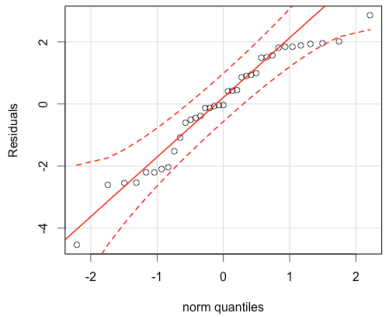


e. CL-SPRT37

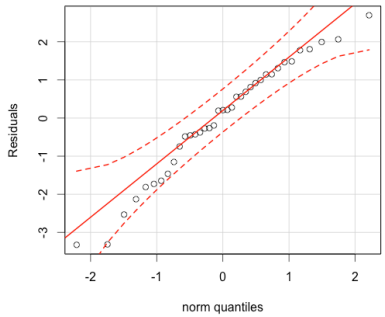


Appendix O. Normality assumption check for CS37 models. QQ plots show model residuals plotted along solid diagonal red line. Normality is violated when excessive numbers of residuals 95% confidence intervals, identified by the red diagonal dotted lines.

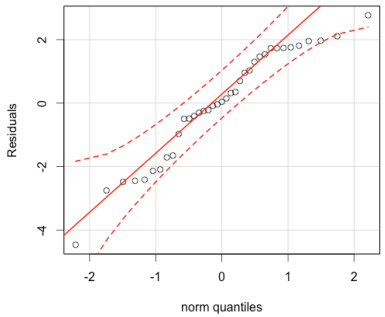
a. DEM37



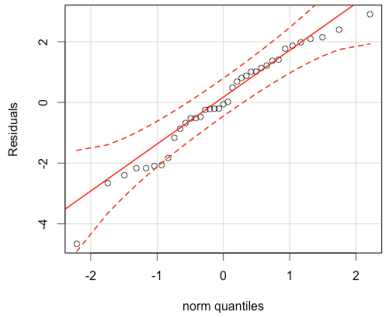
b. PERS37



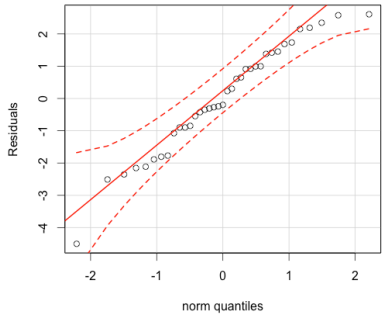
c. ATT37



d. ALL-SPRT37

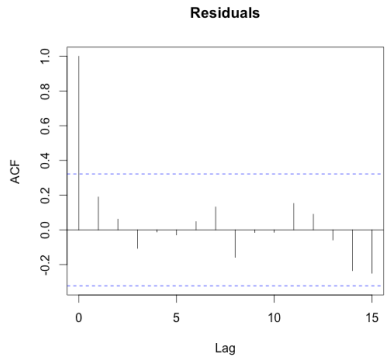


e. CL-SPRT37

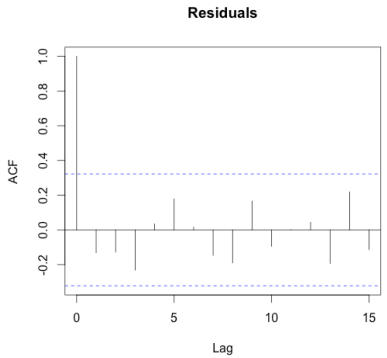


Appendix P. Autocorrelation estimates of CS37 models. Autocorrelation assumptions are met when lags over 0, plotted along x-axis, do not exceed height of dotted horizontal line.

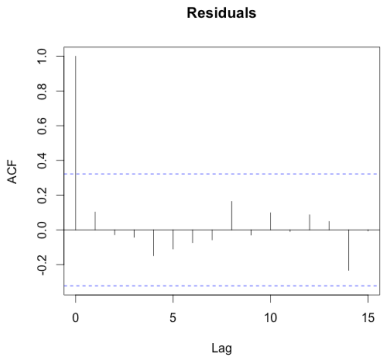
a. DEM37



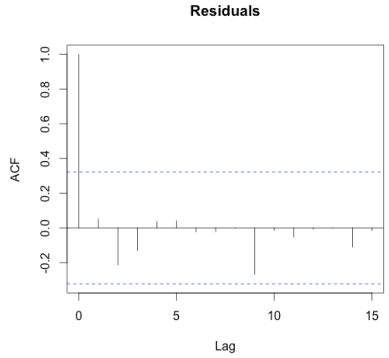
b. PERS37



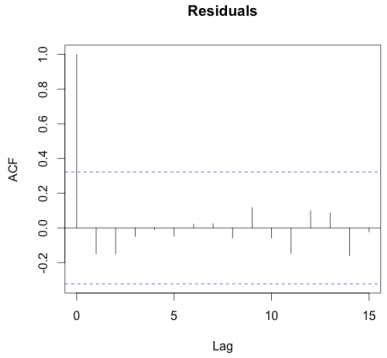
c. ATT37



d. ALL-SPRT37



e. CL-SPRT37



Appendix R. Number of participants who played “Other” sports

Sport	Participants
Cycling	1
Gridiron	1
Kayaking	1
Martial Arts	3
Oz Tag	1
Surf Boat Rowing	1
Surfing	4
Triathlon	2
Wakeboarding	1
Waterskiing	1

Appendix R. Number of participants who played “Other” sports

Sport	Participants
Cycling	1
Gridiron	1
Kayaking	1
Martial Arts	3
Oz Tag	1
Surf Boat Rowing	1
Surfing	4
Triathlon	2
Wakeboarding	1
Waterskiing	1

References

Baker, J., Cote, J., & Abernethy, B. (2003). Sport-specific practice and the development of expert decision-making in team ball sports. *Journal of Applied Sport Psychology*, *15*(1), 12–25.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*(1), 1–26.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *The Journal of Applied Psychology*, *81*(3), 261–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8690688

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and Performance at the Beginning of the New Millennium : What Do We Know and Where Do We Go Next ?, *9*(June).

Batt, L. S., Batt, M. S., Baguley, J. A., & McGreevy, P. D. (2009). The Value of Puppy Raisers' Assessments of Potential Guide Dogs' Behavioral Tendencies and Ability to Graduate. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals*, *22*(1), 71–76.

Bennett, P. C., & Rohlf, V. I. (2007). Owner-companion dog interactions : Relationships between demographic variables , potentially problematic behaviours , training engagement and shared activities, *102*, 65–84. doi:10.1016/j.applanim.2006.03.009

Block, J. (1961). The Q-sort method in personality assessment and psychiatric research.

Bridge, M. W., & Toms, M. R. (2013). The specialising or sampling debate: a retrospective analysis of adolescent sports participation in the UK. *Journal of Sports Sciences*, *31*(1), 87–96.

Budaev, S. V. (2010). Using principal components and factor analysis in animal behaviour research: caveats and guidelines. *Ethology*, *116*(5), 472–480.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press.

Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*(2), 225–248.

De Meester, R. H., Pluijmakers, J., Vermeire, S., & Laevens, H. (2011). The use of the socially acceptable behavior test in the study of temperament of dogs. *Journal of Veterinary Behavior: Clinical Applications and Research*, *6*(4), 211–224.

Detector Dog Program - Working with Dogs to Help Protect Australia. (2012). *Detector Dog Program*. Retrieved April 13, 2014, from http://www.customs.gov.au/site/page4241.asp

Dunbar, I. (2001). *After You Get Your Puppy*. James & Kenneth Publishers. Retrieved from http://books.google.com/books?id=L1KqPQAACAAJ

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., … the R Core Team. (2014). caret: Classification and Regression Training. Retrieved from http://cran.r-project.org/package=caret

Furton, K. G., Hong, Y.-C., Hsu, Y.-L., Luo, T., Rose, S., & Walton, J. (2002). Identification of odor signature chemicals in cocaine using solid-phase microextraction-gas chromatography and detector-dog response to isolated compounds spiked on U.S. paper currency. *Journal of Chromatographic Science*, *40*(3), 147–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11954652

Furton, K. G., & Myers, L. J. (2001). The scientific foundation and efficacy of the use of canines as chemical detectors for explosives. *Talanta*, *54*(3), 487–500. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18968273

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*, 7–28.

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40*(4), 347–358. doi:10.1016/j.jrp.2006.01.001

Goldberg, L. R., Johnson, J. a., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. doi:10.1016/j.jrp.2005.08.007

Gosling, S. D. (1998). Personality dimensions in spotted hyenas (< em> Crocuta crocuta). *Journal of Comparative Psychology*, *112*(2), 107.

Gosling, S. D., Kwan, V. S., & John, O. P. (2003). A dog's got personality: a cross-species comparative approach to personality judgments in dogs and humans. *Journal of Personality and Social Psychology*, *85*(6), 1161.

Gosling, S. D., & Vazire, S. (2002). Are we barking up the right tree? Evaluating a comparative approach to personality. *Journal of Research in Personality*, *36*(6), 607–614.

Gwet, K. (2001). *Handbook of inter-rater reliability* (pp. 152–156). Gaithersburg, MD: STATAXIS Publishing Company.

Harper, R. J., Almirall, J. R., & Furton, K. G. (2005). Identification of dominant odor chemicals emanating from explosives for use in developing optimal training aid combinations and mimics for canine detection. *Talanta*, *67*(2), 313–27. doi:10.1016/j.talanta.2005.05.019

Haverbeke, A., De Smet, A., Depiereux, E., Giffroy, J.-M., & Diederich, C. (2009). Assessing undesired aggression in military working dogs. *Applied Animal Behaviour Science*, *117*(1), 55–62.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586.

Helton, W. S. (2009). *Canine ergonomics: the science of working dogs*. CRC Press.

Hewson, C. J., Luescher, U. A., & Ball, R. O. (1998). Measuring change in the behavioural severity of canine compulsive disorder: the construct validity of categories of change derived from two rating scales. *Applied Animal Behaviour Science*, *60*(1), 55–68.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, *92*(5), 1270.

Hsu, Y., & Serpell, J. A. (2003). Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *Journal of the American Veterinary Medical Association*, *223*(9), 1293–1300.

Jacobs, R., Hofmann, D. A., & Kriska, S. D. (1990). Performance and seniority. *Human Performance*, *3*(2), 107–121.

Jagoe, A., & Serpell, J. (1996). Owner characteristics and interactions and the prevalence of canine behaviour problems. *Applied Animal Behaviour Science*, *47*(1), 31–42.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research*, *3*, 114–158.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Jones, A. C. (2008). *Development and validation of a dog personality questionnaire*. ProQuest.

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

Jones, A. C., Sinn, D. L., Gosling, S. D., & Hilliard, S. (2011). Military working dog field-analysis questionnaire: An evaluation of MWD training and real-world utilization.

Jones, A. C., Sinn, D. L., Gosling, S. D., & Hilliard, S. (2012). Military Working Dog Field-Analysis Questionnaire: Handler evaluations of potentially valuable and potentially detrimental MWD behavioral traits.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*.

Kobelt, A. J., Hemsworth, P. H., Barnett, J. L., & Coleman, G. J. (2003). A survey of dog ownership in suburban Australia—conditions and behaviour problems. *Applied Animal Behaviour Science*, *82*(2), 137–148.

Kotrschal, K., Schöberl, I., Bauer, B., Thibeaut, A.-M., & Wedl, M. (2009). Dyadic relationships and operational performance of male and female owners and their male dogs. *Behavioural Processes*, *81*(3), 383–391.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models.

Ley, J. M., McGreevy, P., & Bennett, P. C. (2009). Inter-rater and test–retest reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R). *Applied Animal Behaviour Science*, *119*(1), 85–90.

Li, A., & Bagger, J. (2006). Using the BIDR to Distinguish the Effects of Impression Management and Self-Deception on the Criterion Validity of Personality Measures: A Meta-Analysis. *International Journal of Selection and Assessment*, *14*(2), 131–141. doi:10.1111/j.1468-2389.2006.00339.x

Maejima, M., Inoue-Murayama, M., Tonosaki, K., Matsuura, N., Kato, S., Saito, Y., … Ito, S. (2007). Traits and genotypes may predict the successful training of drug detection dogs. *Applied Animal Behaviour Science*, *107*(3), 287–298.

McEvoy, G. M., & Cascio, W. F. (1989). Cumulative evidence of the relationship between employee age and job performance. *Journal of Applied Psychology*, *74*(1), 11.

Munsell, K. L., Canfield, M., Templer, D. I., Tangan, K., & Arikawa, H. (2004). Modification of the pet attitude scale. *Society and Animals*, *12*(2), 137–142.

Ng, T. W. H., & Feldman, D. C. (2008). The relationship of age to ten dimensions of job performance. *Journal of Applied Psychology*, *93*(2), 392.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*(2-3), 245–269.

Paul, A. M. (2010). *The cult of personality testing: How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstand ourselves*. Simon and Schuster.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego: Academic Press.

Paunonen, S. V, & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology*, *103*(1), 158.

Penney, L. M., David, E., & Witt, L. a. (2011). A review of personality and performance: Identifying boundaries, contingencies, and future research directions. *Human Resource Management Review*, *21*(4), 297–310. doi:10.1016/j.hrmr.2010.10.005

Podberscek, A. L., & Gosling, S. D. (2000). Personality research on pets and their owners: Conceptual issues and review. *Companion Animals and Us: Exploring the Relationships between People and Pets*, 143–167.

Preylo, B. D., & Arikawa, H. (2008). Comparison of vegetarians and non-vegetarians on pet attitude and empathy. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals*, *21*(4), 387–395.

Rooney, N. J., & Bradshaw, J. W. S. (2002). An experimental study of the effects of play upon the dog–human relationship. *Applied Animal Behaviour Science*, *75*(2), 161–176.

Rooney, N. J., Gaines, S. A., Bradshaw, J. W. S., & Penman, S. (2007). Validation of a method for assessing the ability of trainee specialist search dogs. *Applied Animal Behaviour Science*, *103*(1), 90–104.

Scott, J. P. (1958). Critical periods in the development of social behavior in puppies. *Psychosomatic Medicine*, *20*(1), 42–54.

Scott, J. P. (1965). *Genetics and the Social Behavior of the Dog*. University of Chicago Press.

Serpell, J. a. (1996). Evidence for an association between pet behavior and owner attachment levels. *Applied Animal Behaviour Science*, *47*(1-2), 49–60. doi:10.1016/0168-1591(95)01010-6

Serpell, J., & Jagoe, J. A. (1995). Early experience and the development of behaviour. *The Domestic Dog: Its Evolution, Behaviour, and Interactions with People*, 79–102.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

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

Sinn, D. L., Hixon, G., & Gosling, S. D. (2011). *Exploratory factor analysis, internal validity of aggregate behavior scales, and test-retest correlations in the TSA-CBDC behavioral test data*. (Tasks 2.2, 2.3, and 3.2 Combined Report No. Contract HSHQDC-10-C-00085) (p. 72). Department of Homeland Security, Science & Technology Directorate.

Slabbert, J., & Odendaal, J. (1999). Early prediction of adult police dog efficiency—a longitudinal study. *Applied Animal Behaviour Science*, *64*(4), 269–288.

Sturman, M. (2003). Searching for the Inverted U-Shaped Relationship Between Time and Performance: Meta-Analyses of the Experience/Performance, Tenure/Performance, and Age/Performance Relationships. *Journal of Management*, *29*(5), 609–640. doi:10.1016/S0149-2063(03)00028-X

Svartberg, K. (2002). Shyness–boldness predicts performance in working dogs. *Applied Animal Behaviour Science*, *79*(2), 157–174.

Svartberg, K. (2007). Individual differences in behaviour - Dog personality. In P. Jensen (Ed.), *The Behavioural Biology of Dogs* (pp. 182 – 206). Cambridge: CAB International.

Templer, D. I. (1981). The construction of a Pet Attitude Scale. *The Psychological Record*.

Topál, J., Miklósi, A., & Csányi, V. (1997). Dog-human relationship affects problem solving behavior in the dog. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals*, *10*(4), 214–224.

Van den Berg, L., Schilder, M. B. H., De Vries, H., Leegwater, P. A. J., & van Oost, B. A. (2006). Phenotyping of aggressive behavior in golden retriever dogs with a questionnaire. *Behavior Genetics*, *36*(6), 882–902.

Vazire, S., Gosling, S. D., Dickey, A. S., & Schapiro, S. J. (2007). Measuring personality in nonhuman animals. *Handbook of Research Methods in Personality Psychology*, 190–206.

Whittingham, M. J., Stephens, P. a, Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *The Journal of Animal Ecology*, *75*(5), 1182–9. doi:10.1111/j.1365-2656.2006.01141.x

Wilsson, E., & Sundgren, P.-E. (1997). behaviour test for the selection of dogs for service and breeding, I: Method of testing and evaluating test results in the adult dog, demands on different kinds of service. *Applied Animal Behaviour Science*, *53*(4), 279–295. Retrieved from http://www.sciencedirect.com/science/article/pii/S0168159196011744