Norwegian University
of Life Sciences

Master's Thesis 2019 60 ECTS
Faculty of Biosciences

The role of age in the selection of police patrol dogs using a standardized behavior test

Kim I. Bjørnson
Animal Science, Ethology

Acknowledgments

This master thesis marks my end at NMBU. The last two years have been amazing, and I am thankful for both educational and social experiences. Animal behavior, animal welfare, and statistics have been of great interest to me, and this master thesis gave me an opportunity to work with all three subjects. The process of making this master thesis would not have been possible without the contribution and help from several people, and I would like to express my gratitude to the following:

First, I would like to thank my amazing supervisor Ruth C. Newberry for guidance and constructive feedback throughout the whole project. Your enthusiasm has been inspiring. A special thanks to Judit Vas and Christine Olsen for additional input and encouragement.

I would also like to express my gratitude to all the Norwegian and Swedish police dog testers who participated in this project, as well as all the dog owners for agreeing to bring their dog twice for evaluation.

I also wish to thank my fellow students at IHA for unique and colorful lunch conversations. An extra-large thanks to Johanna Gjøen for always making me laugh. I look forward to sharing an office with you in the future. A special thanks to Komma, who served as a therapy dog during long hours in the study hall. Your cuddles provided a much-needed stress relief, and I will miss you and your blue ball.

I am very grateful to my friends in Ski. Thank you for all your help and support.

Finally, I would like to thank my amazing mom for unconditional support, as well as many hours of dog sitting. Last, but not least, to my loyal dog, Trym, thank you for making sure that my life never gets boring.

University of Life Sciences, Ås

15.05.2019

Kim I. Bjørnson

Abstract

Police patrol dogs face many challenging situations, and only a proportion of dogs are suitable for this work. It is desirable to identify suitable dogs as early as possible, allowing unsuited dogs to be released for other purposes and reducing the emotional cost of separating the dog and handler at a later age. However, the selection process may be less reliable when dogs are young and their personality is less established. I investigated the stability of dog behavior in the different successive subtests of a standardized behavior test conducted at two different ages, and test outcome of each test (pass or fail, based on expert evaluation by testers). I also examined the extent to which behavior in the first test predicted the outcome of the second test.

A standardized test was administered twice to 62 male German shepherds by Norwegian and Swedish police dog testers (N = 31 dogs per country) approximately 6 and 12 months of age (mean ± SD: 6.14 ± 0.50 vs 12.31 ± 0.64 months). Tests comprised 63 behavioral variables assessed across 14 subtests designed to measure behavioral responses in different situations. Each variable was scored from 1 to 5, with higher scores representing more desirable responses, and the mean score for each subtest was calculated.

A positive association was found between test outcomes at 6 and 12 months (χ^2 = 14.78, p <0.001), with 74.2% of dogs having the same outcome at both ages. Bland-Altman plots identified 7 subtests with mean scores that showed consistency in the interval 6-12 months. Binary logistic regression models identified that the mean scores from 3 subtests at 6 months, and 4 at 12 months, were significant predictors of test outcomes at the age tested. Furthermore, 3 subtests at 6 months were significant predictors of test outcomes at 12 months. I compared the mean score from the 3 subtests between dogs that A) passed at both ages (n = 21), B) failed at 6, but passed at 12 months (n = 13) or C) failed at both ages (n = 25). Back-transformed least squares mean scores (± SD) adjusted for multiple comparisons were higher for dogs in category A (4.23 ± 0.36) than B (3.90 ± 0.37, z = 2.62, p = 0.024) or C (3.41 ± 0.35, z = 7.74, p <0.001), and category B scores also exceeded category C scores (z = 4.00, p <0.001).

These results suggest that some subtests are more predictive of test outcomes than others. They also suggest that testing can be implemented at the earlier age to exclude low scoring dogs and accept high scoring dogs while leaving open the possibility of a second test when older for a relatively small subset of young dogs with ambiguous (intermediate) test results.

Sammendrag

Politiets patruljehunder må takle mange utfordrende situasjoner, og bare et fåtall hunder er egnet for dette arbeidet. Muligheten til å identifisere egnete hunder som tidlig som mulig er svært ønskelig, slik at uegnede hunder kan frigjøres til andre oppgaver eller omplasseres. Tidligere omplassering vil redusere den emosjonellbelastningen det er å skille hund og hundefører ved et senere tidspunkt. Seleksjonsprosessen kan være mindre pålitelig når hundene er yngre og deres personlighet er mindre etablert. Jeg undersøkte stabiliteten til hundeadferd i ulike suksessive deltester i en standardisert adferdstest utført ved to ulike aldere, og testresultatet fra hver testalder (bestått vs. stryk, basert på ekspertevaluering fra testledere). Jeg undersøkte også til hvilken grad adferden i den første testen predikterte testresultatet i den andre testen.

I dette prosjektet ble 62 Schäferhund hanner testet av norske og svenske testere (N = 31 hunder fra hvert land) når de var 6 og 12 måneder gamle (gjennomsnittsalder ± standardavvik: 6.14 ± 0.50 vs. 12.31 ± 0.64 måneder). Testen besto av 63 variabler vurdert over 14 deltester designet for å måle adferd i ulike situasjoner. Hver variabel ble gitt en verdi fra 1-5, og gjennomsnittsverdier ble regnet ut for alle deltestene.

En positiv assosiasjon ble funnet mellom testresultatene ved 6 og 12 måneders alder ($\chi^2 = 14.78$, $p < 0.001$), hvor 74.2% av hundene fikk samme resultat ved begge testene. Bland-Altman figurer identifiserte 7 deltester med gjennomsnittsverdier som var stabile i intervallet 6-12 måneder. Binære logistiske regresjonsmodeller fant at gjennomsnittsverdier fra 3 deltester ved 6 måneder, og 4 ved 12 måneder, var signifikante prediktorer for testresultatet ved de to alderne. Jeg fant også at 3 deltester ved 6 måneder var signifikante prediktorer for testresultat ved 12 måneders alder. Jeg sammenlignet gjennomsnittsverdien av de 3 deltestene mellom A) hunder som besto begge testene (n = 21), B) hunder som strøk ved 6, men besto ved 12 måneders alder (n = 13), og C) hunder som strøk på begge testene (n = 25). Tilbake-transformert gjennomsnitt minstekvadrat verdier (least squares mean scores) justert for flere sammenligninger var høyere for hunder i gruppe A (4.23 ± 0.36) enn gruppe B (3.90 ± 0.37, z = 2.62, $p = 0.024$) og gruppe C (3.41 ± 0.35, z = 7.74, $p < 0.001$). Gruppe B hadde også høyere verdier enn gruppe C (z = 4.00, $p < 0.001$).

Disse resultatene antyder at noen deltester er mer prediktive av testresultat enn andre. Resultatene antyder også at testing kan bli iverksatt ved en tidligere alder for å utelukke hunder med lave verdier og akseptere hunder med høye verdier, med mulighet for å teste unge hunder med tvetydige (uklare) testverdier på nytt når de blir eldre.

Table of Content

1. Introduction

1.1. Background

Domestic dogs (*Canis lupus familiaris*) have many roles in the modern human society, ranging from being loyal companions to providing crucial assistance, as working dogs, a term used in this paper for police and military dogs, and service dogs (e.g. guide dogs). A common problem is that many working and service dogs never successfully complete training and enter active service (Cobb et al., 2015). Slabbert and Odendaal (1999) reported that 70 % of the dogs bred at the South African Police Dog Breeding Center (SAPSDBC) were rejected as police dogs. Similarly, only 27 % of the dogs from the Swedish Armed Forces (SAF) breeding program were deemed suitable as police or military dogs between the start of the program in 2005 and 2010 (Foyer et al., 2013). Dogs selected as working dogs have personalities that differ from the general population, making them suitable for a specific working role (Wilsson & Sundgren, 1997), and most dogs are rejected because they exhibit unsuitable behavior (Duffy & Serpell, 2012; Foyer et al., 2013; Slabbert & Odendaal, 1999). Identifying dogs with the desired personality as early as possible has been the focus of several studies over the past decades (e.g. Goddard & Beilharz, 1986; Harvey et al., 2016b; Wilsson & Sundgren, 1997), and successful assessment of such traits can reduce both the time and financial cost of rearing and training a potential working dog, as well as ensuring that rejected dogs can be re-homed as early as possible.

1.2. Personality of working dogs

Working dogs experience a variety of stressful and demanding situations, and correctly determining which dogs are capable of such work is important. Not only is the performance of working dogs correlated with their personality (Hoummady et al., 2016; Sinn et al., 2010; Svartberg, 2002), but placing an unqualified dog in active service could have serious consequences. For example, a unsuited police dog might fail to provide assistance when needed, or it may react badly (e.g. aggressively) when exposed to aversive or startling situations (e.g. being threatened or exposed to loud noises), potentially ending up posing a danger to its handler or civilians (Slabbert & Odendaal, 1999). Determining which dogs will react appropriately in different situations is important when selecting working dogs.

One definition of personality or temperament is "the underlying behavioral tendencies that differ across individuals, that are consistent within individuals over time, and that affect the behavior that is expressed in different contexts" (Stamp & Groothuis, 2010, p. 302). Five personality traits – sociability, playfulness, chase-proneness, aggressiveness, and curiosity/fearlessness – and a broad personality dimension, the shyness-boldness dimension, have been suggested in dogs (Svartberg & Forkman, 2002). These personality traits, with the exception of aggressiveness, have been related to the shyness-boldness dimension (Svartberg & Forkman, 2002), which is correlated with working dog performance, specifically with bolder dogs having better test performance (Svartberg, 2002) in the mentality assessment (DMA), a personality test originally designed to assess personality for breeding and selection of working dogs (Svartberg, 2002; Svartberg & Forkman, 2002; Svartberg et al., 2005).

1.3. Selection and qualification of working dogs

Selective breeding of dogs originally started to improve work performance in task such as hunting, guarding and herding (van den Berg, 2017). Over the centuries, selective breeding has not only resulted in the more than the 400 different dog breeds we recognize today (Careau et al., 2010; Jamieson et al., 2017), but also a number of working breeds that are highly specialized for specific tasks (Lord et al., 2014; Lord et al., 2017). Today, selection of working dogs takes place in many different ways. Some police and military agencies have established their own breeding programs (e.g. SAF), but the majority rely on private vendors or breeders to purchase dogs (Rooney et al., 2016). Many dogs are purchased as puppies (8 weeks old), and then placed either with their future handler or with a volunteer sometimes referred to as a 'puppy raiser' (Wilsson & Sinn, 2012) or 'puppy walker', though it is not uncommon to obtain adult dogs (>1 years old) from private vendors (Sinn et al., 2010). One commonality between most working dog programs is a qualification test which the dogs must pass before they can continue with further training (e.g. Sinn et al., 2010; Slabbert & Odendaal, 1999; Wilsson & Sundgren, 1997).

Qualification of working dogs is assessed by a standardized behavioral test (also called 'temperament', 'mentality', or 'personality' test). The specific layout of the test varies somewhat between different programs and agencies, but generally consists of a series of subtests designed to assess dogs' behavioral responses in situations simulation those they might encounter in active service. Subtests commonly present in working dog qualification tests measure a dog's behavior when in contact with people, environmental sureness, focus and determination during search, gun sureness, interest in play, and the tendency to defend

itself and handler, as well as the ability to overcome and recover from fearful or aversive stimuli (Sinn et al., 2010; Svartberg & Forkman, 2002; Svartberg, 2005; Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012). If the dogs pass the qualification test, they enter a training program. Previous studies have found that the likelihood of completing training is associated with the dog's tendency to defend its handler or itself, willingness to participate in competitive games (e.g. tug-of-war) and chase moving objects, and the ability to overcome and recover from fearful and stressful situations (Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012). Therefore, it is logical to assess these characteristics in the qualification test.

1.4. Predicting future behavior

In behavioral testing, there are two important concepts; validity and reliability (Diederich & Giffroy, 2006; Taylor & Mills, 2006). Validity refers to how well a variable (e.g. a behavioral measurement) actually measures what it is supposed to measure. More specifically, validity is an indicator of the association between the measured behavioral variable and what the variable is meant to predict (Martin & Bateson, 2007). There are various of ways to evaluate validity (see Tylor & Mills, 2006), and one of these – predictive validity – is especially important when trying to assess the future behavior of an individual (Sinn et al., 2010). Predictive validity describes how well a behavioral measurement (e.g. behavior score or test outcome) predicts later performance (Diederich & Giffroy, 2006; Taylor & Mills, 2006), such as working dogs passing a qualification test or completing training. Reliability measures the degree of which behavioral measurements are free from random errors. It describes the repeatability and consistency of a measurement (Martin & Bateson, 2007). One important assessment is test-retest reliability, which measures the consistency within the dog itself (Taylor & Mills, 2006).

Consistency is one of the criteria for personality. However, this stands in contrast to the expression 'personality development', which suggests that the expression of behavior in different situations may change during an animal's lifetime. A better term when measuring personality is temporal consistency, which refers to behavioral patterns or tendencies being consistent over a period of time (Stamps & Groothuis, 2010). Knowledge about personality consistency is especially important when we want to predict future behavior based on a single behavioral test (Svartberg et al., 2005).

There is evidence that personality consistency in dogs is affected by age (Fratkin et al., 2013; Goddard & Beilharz, 1986), and a meta-analysis found personality consistency to be significantly higher in dogs older than 12 months (mean $r = 0.51$) compared to dogs younger

than 12 months (mean r = 0.31) (Fratkin et al., 2013). One of the factors suggested to have an effect on personality consistency is the animal's age of maturation (Stamps & Groothuis, 2010; Svartberg et al., 2005). Dogs typically reach sexual maturity between 6 and 9 months of age, but most dogs do not reach social, or behavioral, maturity before 12 and 24 months of age, depending on the breed (Overall, 2013). This might explain why most dogs do not exhibit fully adult behavior until around 2 years of age (Miklósi, 2015), and why testing puppies (~8 weeks) to assess adult behavior might give little information. With the exception of some studies (Slabbert & Odendaal, 1999; Svobodová et al., 2008), puppy tests are generally reported to provide low to no predictability of adult behavior (Goddard & Beilharz, 1986; Riemer et al., 2014; Wilsson & Sundgren, 1998). The stability of personality traits attributed to puppies is largely affected by internal and external changes occurring during development (Miklósi, 2015), which in turn affect the predictive validity of puppy tests. The predictive validity will decrease if the test-retest reliability (i.e. consistency) is low (Patronek et al., 2019).

The juvenile period, which is usually defined to last from approximately 12 weeks (end of the socialization period) until the dog reaches sexual maturity (Serpell et al., 2017), is one of the least studied periods in dogs (Miklósi, 2015). However, there is evidence that evaluation of dog personality traits as early as 5 months of age is somewhat predictive of adult behavior in guide dogs (Harvey et al., 2016b; Serpell & Duffy, 2016). This suggests that is possible to increase predictive validity by testing juvenile dogs rather than puppies.

Potential police dogs are usually subjected to a qualification test at approximately 1-1.5 years of age (e.g. Wilsson & Sinn, 2012), and desirable personality traits vary somewhat between specific working roles (Goold et al., 2016). Police detection dogs search for contraband (e.g. drugs or money) (Goold et al., 2016), and motivation to search is especially desirable (Jamieson et al., 2017). Police patrol dogs perform a range of different tasks (e.g. detaining a suspect, patrolling the streets, and controlling large crowds) (Goold et al., 2016), and suitable dogs are selected based on several personality traits (Wilsson & Sundgren, 1997). There are, to my knowledge, no studies to date investigating the predictive validity of tests on juvenile police patrol dogs.

1.5. Aim, hypotheses and predictions

The main aim of this study was to investigate if the qualification test for Norwegian and Swedish police patrol dogs could be conducted at 6 months of age instead of the present standard of 12 months of age in Norway and 18 months of age in Sweden. Additionally, I

wanted to assess which behavioral responses (i.e. used as measures of personality) were associated with test outcome (pass vs fail), as well as the consistency of these measures over time. Lastly, I wished to see if it was possible to detect which dogs would pass the qualification test at 12 months, despite having failed at 6 months of age (i.e. predict future improvement). To achieve these goals, I assessed the (1) predictive validity by comparing the test outcome at two test ages, (2) test-retest reliability (i.e. temporal consistency) between subtests at two different ages, (3) predictive value of subtests, and (4) difference in test performance between dogs that achieved the same test outcome at the two ages and dogs that failed at 6 months, but passed at 12 months of age.

I hypothesized that testing potential police dogs when they are 6 months of age can provide important and representative insight into their qualification test results at 12 months of age. I expected that, if this hypothesis (i.e. predictive validity) is true, there would be a strong association between the test outcome (i.e. pass vs fail) at the two test ages, with most dogs receiving the same test outcome at both test ages. Because results from previous studies show that not all subtests are equally associated with the test outcome (e.g. Harvey et al., 2016b; Wilsson & Sinn, 2012), I expected to see such differences in this study as well. Furthermore, since predictive validity and reliability are correlated (Patronek et al., 2019), I expected to find temporal consistency between subtest behavior scores at 6 and 12 months, among those subtests conducted at 6 months that had scores associated with test outcome at 12 months. In subtests with low temporal consistency, I expected the dogs to have higher subtest behavior scores at 12 than 6 months of age, indication an improvement in suitability for police work at the higher age.

I also hypothesized that the degree of future improvement in suitability can be detected based on differences in test performance at 6 months between dogs with different test outcomes. If so, I expected a difference in behavior scores depending on the test outcome, with dogs that passed having a higher overall test score (summed over scores for behavior during subtests) than dogs that failed. Moreover, I expected that dogs that passed at 12 months after failing at 6 months of age, would have higher scores at 6 months compared to dogs that failed at both test ages. An overview of the hypotheses and predictions is listed in Table 1.

Table 1. Hypotheses (H) and corresponding predictions (a-c).

Hypotheses and predictions	
H1.	Testing potential police dogs at 6 months of age provides predictive validity for test results at 12 months of age.
a.	There is a strong association between test outcome (pass vs fail) at 6 months and 12 months, whereby most dogs receives the same test outcome at both test ages.
b.	In subtests with behavior scores at 6 months associated with test outcome at 12 months, there is temporal consistency in the behavior scores across the 6-12 months interval.
c.	In subtests with low temporal consistency, I expect the dogs to perform better (i.e. score higher) at 12 months of age
H2.	There is a difference in subtest behavior scores between dogs depending on the test outcome.
a.	Passing dogs have a higher subtest behavior scores than dogs that fail, at both 6 and 12 months of age.
b.	Dogs that fail at 6 months but pass at 12 months, have higher behavior scores at 6 months than dogs that fail at both 6 and 12 months.

2. Material and methods

2.1. Ethical considerations

The data were collected internally by the Norwegian and Swedish police, and I received anonymous data with no personal information about humans or dogs. The project involved no invasive methods on dogs. Dog keepers were informed about the study and consented to bring their dog for two tests. They were free to withdraw from the study at any time. Only the 12-month test result was used to decide whether to accept the dog for police dog training.

2.2. Subjects

The subjects of this study were male German shepherds in Norway and Sweden that were tested once at approximately 6 months and again at approximately 12 months of age (mean ± SD: 6.14 ± 0.50 vs 12.31 ± 0.64 months). The initial sample size consisted of 75 dogs; 41 and 34 Norwegian and Swedish dogs, respectively. Dogs that were only tested once (at 6 months) were not included in the analyses, reducing the final sample size to 62 dogs; 31 from each country. Only one neutered dog participated in the study, the rest being intact.

Subjects were chosen because they are the most representative subjects for the Norwegian and Swedish police dog population. German shepherd is the most common breed used for police patrol dogs (Goold et al., 2016). Similarly, males are more commonly used as patrol dogs than females (Goold et al., 2016; Sinn et al., 2010), presumably because males are bolder than females (Svartberg, 2002), which, combined with their larger size, makes them more suitable for work in law enforcement (Svobodová et al., 2008).

2.3. Test procedure and behavioral rating

The dogs were tested using a standardized behavioral test designed to measure the dogs' reaction to various situations they might encounter during active service (e.g. sudden noises, threatening figures, and gunshots). The behavioral test used in this study was adapted from already existing test procedures to harmonise methods between the two countries, and was similar to tests such as DMA used in other studies and working dog programs (e.g. Svartberg & Forkman, 2002; Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012). The behavioral test was identical at both test ages. All Norwegian tests took place at Hauerseter, NO, and all the Swedish test were conducted at Karlsborg, SE. The data were collected from 3. November 2015 to 11. June 2018, and the tests were always carried out between 7.30 AM and 4.00 PM.

The behavioral test consisted of 14 subtests, mainly carried out outside except where stated otherwise, and always in the same order (Section 2.4). The average test time was 69 minutes (± 20.21 SD, N = 62), and the dogs were given no breaks between subtests apart from the time it took to move between stations. The dogs were tested individually. Each dog was accompanied throughout the test by a dog keeper (owner or another familiar person), and usually two trained testers, who conducted the test and guided the dog keeper on how to act during the test. During testing, the dog keeper or a tester served as the dog's handler and another tester controlled or created stimuli used in subtests. To obtain data for all subtests, the test was completed even if it became evident before completion that the dog would fail. However, the test was stopped at any point on the dog keeper's request or if the dog showed signs of too much stress. (e.g. unwilling to move to the next station, highly fearful or aggressive, panting heavily).

There were 63 behavioral variables assessed across the 14 subtests, and each variable was scored from 1 to 5, unless stated otherwise. The scores were mutually exclusive, and higher scores represented more desired responses (e.g. aggression: 5 = "relaxed", 1 = "alert, tries to bite"). A standardized score sheet with a behavioral description for each score level was provided to ensure that scoring was as objective as possible. Testers from both countries met before data collection started to practice the test procedures, evaluate and improve scoring consensus, and refine definitions and procedures to maximise inter-tester reliability. There was a total of 14 different testers in the study. Behavior in each subtest was scored before moving to the next subtest. The dog's overall suitability to serve as a police patrol dog was evaluated at the end of the test, and each dog was scored as passed or failed based on consensus between testers.

2.4. The subtests

2.4.1. Social contact

This subtest, conducted outdoors, measured the dog's reaction to strangers. The handler walked the dog on a leash toward a group of strangers standing passively in a line. The dog was walked passed the group in close proximity (<1 meter), and the strangers did not interact with the dog. The handler then led the dog away from the group, and a tester approached the dog. After greeting the dog, the tester took the leash and led the dog away from the handler to perform a physical examination. The tester touched the dog's sides, back, around the mouth, and hind legs. The dog's response was measured by the following five behavioral variables.

Contact with strangers: Scored from "initiates contact with strangers" (5), to "rejects strangers with aggression" (1). Described when the dog was walked passed the group.

Social confidence: Scored from "balanced greeting behavior" (5), to "rejects with aggression" (1). Described when the dog was walked passed the group.

Contact with tester: Scored from "initiates contact with tester" (5), to "rejects tester with aggression" (1). Described when the tester approached the dog.

Following: Scored from "follows willingly" (5), to "does not follow willingly, tester must use the leash to get the dog to follow" (1). Described when the tester led the dog away from the handler.

Handling: Scored from "accepts handling" (5), to "rejects handling with aggression" (1). Described during the physical examination.

Confidence: Scored from "relaxed, tail-wagging, confident posture, interacts with humans" (5), to "flees or backs away from humans, urinating" (1). Described for behavior during the whole subtest.

2.4.2. Playfight

This subtest, conducted outdoors, measured the dog's behavior during tug-of-war. The dog was unleashed, and a tester started playing with a long, strong rag (or tug) to get the dog's interest, before offering the rag to the dog, inviting it to a game of tug-of-war. During the game, the tester increased and decreased the strength of pulling on the rag (i.e. changing the resistance the dog experienced). The dog's response was measured by the following six behavioral variables.

Intensity: Scored from "high intensity" (5), to "very low intensity, shows hesitation or reluctance" (1). Described when, and if, the dog lunged after the offered rag.

Grip strength: Scored from "grips the rag with full bite" (5), to "weak grip, thin bite" (1). Described when, and if, the dog first took the rag.

Drive: Scored from "fights intensely, increases fight with resistance, high drive" (5) to "do not fight, insignificant drive" (1). Described during the tug-of-war.

Resilience: Scored from "highly resilient to resistance" (5), to "reacts fearfully when experiencing resistance" (1). Described during tug-of-war when the tester increased the pull strength.

Aggression: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described for behavior during the whole subtest.

Confidence: Scored from "relaxed, tail-wagging, confident posture, interacts with tester" (5), to "flees or backs away from tester" (1). Described for behavior during the whole subtest.

2.4.3. Retrieval

This subtest measured the dog's willingness to chase after and retrieve a ball (or kong) in an outdoor environment. The dog was unleashed, and the handler threw the ball for the dog to chase after. When, and if, the dog picked up the ball, the handler called the dog back. If the dog did not return to the handler after the first call, the handler would repeat the command. If the dog returned, but did not release the ball, the handler would give a new command to release the ball. The dog's response was measured by the following three behavioral variables.

Cooperation: Scored from "returning with and releasing the ball to handler on first command" (5), to "ignoring the ball" (1). Described after the handler threw the ball.

Aggression: Score from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described for behavior during the whole subtest.

Confidence: Score from "relaxed, tail-wagging, confident posture, interacts with handler" (5), to "flees or backs away from handler" (1). Described for behavior during the whole subtest.

2.4.4. Search outdoors

This subtest tested the dog's ability and motivation to search for and locate a hidden object in an outdoor environment. A tester hid a toy (e.g. ball or kong) in a 25 x 25 meter area outside. The dog was present when the toy was hidden. The dog was unleashed and given a command to start the search. The dog was given verbal encouragement if it showed little interest or was distracted during the search. The search lasted until the dog located the toy, or until the test was stopped due to lack of interest or success. If the dog located the toy, the handler would call the dog back. If the dog did not return, the handler would repeat the command. If the dog returned, but did not release toy, the handler would give a new command to release the toy. The dog's response was measured by the following four behavioral variables.

Time in search: Time (in seconds) from when the dog was unleased until it located the toy.

Focus: Scored from "searches focused and efficiently" (5), to "low interest in searching despite encouragement" (1). Described during the search.

Tracking ability: Scored from "great tracking ability, follows the track" (5), to "small or no interest in the hidden object" (1). Described during the search.

Cooperation: Scored from "returning with and releasing the toy to handler on first command" (5), to "ignores the toy" (1). Described when, and if, the dog located the toy.

2.4.5. Sudden noise

This subtest measured the dog's curiosity and motivation to approach and explore a novel sound in an outdoor environment. The unleashed dog was held by its collar, while a tester stood out of sight and snapped a twig. The dog was then released. The dog's response was measured by the following four behavioral variables.

Reaction to noise: Scored from "reacts quickly, gets excited" (5), to "no reaction" (1). Described when the tester made the noise.

Curiosity: Scored from "runs straight to tester" (5), to "does not approach tester" (1). Described after the dog was released.

Aggression: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described when, and if, the dog approached the tester.

Confidence: Scored from "relaxed, tail-wagging, confident posture, interacts with tester" (5), to "flees or backs away from tester (1). Described when, and if, the dog approached the tester.

2.4.6. Hunting drive

This subtest measured the dog's willingness to run after a moving object in an outdoor environment. The dog was leashed and stood or sat next to the handler or a tester. A tester was crouched down under a canvas approximately 10-15 meters away. The tester started running away from the dog, still hunched over and holding the canvas over their head. The tester varied between running and sitting down a few times, before sitting down and remaining passive under the canvas. The dog was unleashed when the tester started running. If the dog did not approach the tester, either during the run or when the tester sat down, the handler would support the dog by walking over to the tester. The dog's response was measured by the following four behavioral variables.

Intensity: Scored from "executes with high speed" (5), to "will not start" (1). Described when the dog was unleashed.

Interest: Scored from "shows no fear" (5), to "will not approach" (1). Described when the dog was unleashed.

Aggression: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described when, and if, the dog approached the tester.

Confidence: Score from "relaxed, tail-wagging, confident posture, interacts with tester" (5), to "flees or backs away from tester" (1). Described when, and if, the dog approached the tester.

2.4.7. Sudden appearance

This subtest measured the dog's response when a human-like dummy suddenly appeared in front of the dog while walking outdoors. A boiler-suit was used to create a human-like dummy. The dummy's legs were secured to the ground and a rope was fastened to the arms. The rope led up to a wooden bar fastened between two trees or wooden poles. The rope was directed away from the set-up to a tester standing approximately 5 meters away. At the start of the subtest, the dummy lay folded on the ground, not visible to the dog. Pulling on the rope caused the dummy to suddenly appear in a standing position with its arms raised upwards. The handler walked the leashed dog toward the location of the dummy. When the dog came close (1-2 meters), the rope was pulled and the dummy appeared suddenly in front of the dog. The handler immediately let go of the leash, giving the dog room to react (e.g. evasive maneuvers or approach). The handler remained passive for 15-20 seconds, allowing the dog to approach the dummy by itself. If the dog did not approach the dummy, the handler assisted the dog in steps until the last step was executed or the dog approached the dummy.

1. Handler takes 1-2 steps closer to the dummy.
2. Handler approaches the dummy and makes physical contact with it.
3. The dummy is lowered to the ground.

After the dog approached the dummy, or the last step was executed, the handler took the leash and led the dog away. After approximately 5 meters, the handler turned and walked the dog past the dummy. The handler led the dog past the dummy twice. The dog's response was measured by the following four behavioral variables.

Startle response: Scored from "no avoidance, continues to walk" (5), to "runs away" (1). Described when the dummy appeared.

Defense: Scored from “clear defensive reaction” (5), to “no threats” (1). Described when, and right after, the dummy appeared.

Exploration: Scored from “approaches the dummy by itself” (5), to “approaches the dummy with handler when the dummy is lowered to the ground” (1). Described when, and if, the dog approached the dummy.

Avoidance: Scored from “no signs of fear or evasive maneuvers when passing the dummy” (5), to “evasive maneuvers at every passing of the dummy” (1). Described during the repeated passes of the dummy.

2.4.8. Metallic noise

In this subtest, the dog’s reaction to a loud noise was tested while walking in an outdoor environment. The noise was created by letting metal objects drop down a metal ramp. The ramp consisted of a corrugated metal sheet (width ≈ 1 meter, height = 1.5-2 meters), standing vertically supported by a wooden structure. Metal objects (e.g. metal buckets and a metal chain) were held on the top of the ramp by a rope. The handler led the leashed dog on a walk towards the ramp. When the dog was adjacent to the ramp, the tester let go of the rope, and the metal objects fell down corrugated metal surface, creating a loud noise. The handler let go of the leash when the noise occurred, and remain passive for 15-20 seconds afterwards letting the dog investigate the ramp. If the dog did not approach the ramp on its own, the handler assisted the dog in steps until the last step was executed or the dog approached the ramp.

1. Handler takes 1-2 steps closer to the ramp.
2. Handler approaches the ramp.
3. Handler makes physical contact with the ramp.

When the dog approached the ramp, or the last step was executed, the handler led the dog away. After approximately 5 meters, the handler turned and led the dog past the ramp twice. The dog’s response was measured by the following four behavioral variables.

Startle response: Scored from “no evasive maneuvers” (5), to “runs away” (1). Described when the noise occurred.

Exploration: Scored from “approaches the metal directly with full attention” (5), to “does not approach” (1). Described when the dog was unleashed.

Aggression: Scored from “relaxed” (5), to “alert, lunges, tries to bite” (1). Described when, and right after, the noise occurred.

Avoidance: Scored from "no signs of fear or evasive maneuvers when passing the metallic ramp (5), to "evasive maneuvers at every passing of the ramp" (1). Described during the repeated passes of the ramp.

2.4.9. Sled

In this subtest, conducted outdoors, the dog's reaction to the approach of a novel object was measured. The novel object was a miniature sled with a square wooden base, approximately 0.5 x 0.5 meters, with a carboard figure shaped like a human torso (head and upper body) on top. A rope was secured to the front of the sled. At the start of the subtest, a tester, the handler, and the leashed dog stood on a line facing the sled, which was 15-20 meters away. The tester pulled the sled closer in intervals, keeping it still for short intervals before pulling it forward again. When the sled was 3-5 meters away, the tester tugged hard on the rope, causing the sled to shoot abruptly forward. The handler remained passive when the sled was moving, and the dog was allowed to move freely in the leash. The dog was unleashed and the handler remained passive for 15-20 seconds. If the dog did not approach the sled on its own, the handler assisted the dog in steps until the last step was executed or the dog approached the sled.

1. Handler takes 1-2 steps closer to the sled.
2. Handler approaches the sled.
3. Handler makes physical contact with the sled.

When the dog approached the sled, or the last step was executed, the handler led the dog past the sled twice. The dog's response was measured by the following six behavioral variables.

Defense: Scored from "walks towards the sled" (5), to "clear avoidance, submissive behavior (e.g. low posture)" (1). Described when the sled was pulled towards the dog.

Startle response: Scored from "no evasive maneuvers, stands in front of handler" (5), to "runs away" (1). Described when the sled was pulled abruptly towards the dog.

Threat response: Scored from "proportioned aggression" (5), to "excessive aggression when the threat ceases" (1). Described when the sled was moving and after it stopped.

Exploration: Scored from "approaches the sled independently" (5), to "refuses to approach the sled" (1). Described when the sled stopped moving and the dog was unleashed.

Avoidance: Scored from "no signs of fear or evasive maneuvers when passing the sled (5), to "evasive maneuvers at every passing of the sled" (1). Described during the repeated passes of the sled.

Aggression: Scored from “relaxed” (5), to “alert, lunges, tries to bite” (1). Described for behavior during the whole subtest.

2.4.10. Ghost

This subtest measured the dog’s reaction when approached by a threating and masked individual (i.e. ‘ghost’) in an outdoor environment. The ghost was a tester wearing white clothes and a bucket-like mask. The mask was white with eyes and mouth painted in black. The handler stood with the leashed dog facing in the direction where the ghost was hidden behind a tree 15-20 meters away. The subtest started when the ghost moved into the dog’s view and started to approach the dog in an unnatural and threatening manner; bent slightly forward, sneaking towards the dog. The ghost moved in intervals, stopping and staring at hand signals from the tester. The handler remained passive during the approach. When the ghost came close (< 2 meters) it made a sudden jump towards the dog before standing still and remaining passive. The handler then let go of the leach, letting the dog approach. If the dog did not approach the ghost, the handler assisted the dog in steps until the last step was executed or the dog approached the ghost.

1. Handler takes 1-2 steps towards the ghost.
2. Handler approaches the ghost.
3. Handler makes physical contact with the ghost.

The dog’s response was measured by the following five behavioral variables.

Threats: Scored from “makes clear threats, pulls on the leash” (5), to “makes no threats towards the ghost” (1). Described during the approach.

Startle response: Scored from “no evasive maneuvers, stands in front of handler” (5), to “runs away” (1). Described when the ghost jumped towards the dog.

Exploration: Scored from “approaches the ghost independently” (5), to “will no approach the ghost” (1). Described when the ghost stood passive after the jump.

Aggression: Scored from “relaxed” (5), to “alert, lunges, tries to bite” (1). Described when, and if, the dog approached the ghost.

Confidence: Scored from “relaxed, tail-wagging, confident posture, interacts with the ghost” (5), to “flees or backs away from the ghost (1). Described when, and if, the dog approached the ghost.

2.4.11. Environment substrate

This subtest tested the dog's environmental sureness when moving on challenging floor surfaces. The subtest was conducted in two parts. In the first part, the handler walked the leashed dog up and down a set of steep stairs with metal grate steps. This was repeated twice. The distance between the steps increased towards the top of the stairs, which was approximately 3 meters above the ground. In the second part of the subtest, the hander led the dog inside an unfamiliar room. The floor in the room was shiny and slippery. The dog was unleashed and allowed to moving freely around the room. After letting the dog explore the room, the handler took out a rag and tried to engage the dog in a game of tug-of-war on the slippery floor. The dog's confidence on the stairs and in the room was scored 1, 3, or 5, while the dog's aggression was scored normally (i.e. 1-5). The dog's response was measured by the following five behavioral variables.

Aggression 1: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described when the dog was walked up and down the stairs

Confidence 1: Scored from "relaxed, tail-wagging, confident posture, moves and behaves well on the stairs" (5), to "too scared or distracted to move freely in the environment" (1). Described when the dog was walked up and down the stairs**.**

Aggression 2: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described when the dog was unleashed in the unfamiliar room.

Confidence 2: Scored from "relaxed, tail-wagging, confident posture, moves and well on the slippery floor" (5), to "too scared or distracted to move freely in the environment" (1). Described when the dog was unleashed in the unfamiliar room.

Play: Scored from "grips the rag with full bite, fights intensely" (5), to "not engaged in play" (1). Described during the tug-of-war.

2.4.12. Dark environment

This subtest tested the dog's sureness when moving around in a dark indoor environment, and was conducted directly after, and in the same room, as the previous subtest. After the tug-of-war, the rag was removed and the lights were turned off. The tester then threw an object on the floor (e.g. rag or toy), creating a sound to capture the dog's attention towards the object. The dog's confidence in the dark was scored 1, 3, or 5. The other variables were scored on a 1-5 scale as usual. The dog's response was measured by the following three behavioral variables.

Aggression: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described for behavior during the whole subtest.

Confidence: Scored from "relaxed, tail-wagging, confident posture, moves well in the environment" (5), to "too scared or distracted to move freely in the environment (1). Described for behavior during the whole subtest.

Curiosity: Scored from "runs straight to the stimulus" (5), to "will not approach stimulus" (1). Described when the tester threw the object on the floor.

2.4.13. Search indoors

This subtest tested the dog's ability to search and locate a hidden object in a demanding and distracting indoor environment. A tester hid a toy (e.g. ball or kong) inside a cluttered room unfamiliar to the dog. The dog was unleashed upon entering the room, and was given a command to search for the toy. The dog was given verbal encouragement if it lost interest or became distracted by the environment. The search lasted until the dog located the toy, or until the test was stopped due to lack of interest or success. If the dog located the toy, the handler would call the dog back. If the dog did not return, the handler would repeat the command. If the dog returned, but did not release toy, handler would give a new command to release the toy. The dog's response was measured by the following four behavioral variables.

Time in search: Time (in seconds) from when the dog was unleased until it located the toy.

Focus: Scored from "searches focused and efficiently" (5), to "low interest in searching despite encouragement" (1). Described during the search.

Tracking ability: Scored from "great tracking ability, follows the track" (5), to "small or no interest in the hidden object" (1). Described during the search.

Cooperation: Scored from "returning with and releasing the toy to handler on first command" (5), to "ignores the toy" (1). Described when, and if, the dog located the toy.

2.4.14. Gunshot

In this subtest, the dog's reaction to gunshots was measured. The dog was leashed, and the handler took the dog for a short walk outdoors. During the walk, a tester fired two blank gunshots from a handgun, a few seconds apart. After the gunshots were fired, the handler engaged the dog in a game of tug-of-war. The tester fired two new gunshots during the tug-of-war. When the gunshots were fired, the handler continued the activity and did not react to the

sound. The dog was allowed to move freely on the leash, and the handler let go of the leash if the dog needed more room (e.g. wanted to flee). The dog's reaction was measured by the following five behavioral variables.

Reaction: Scored from "no reaction" (5), to "flight tendencies, signs of fear" (1). Described when the gunshots were fired during the walk.

Aggression: Scored from "relaxed" (5), to "alert, lunges, tries to bite" (1). Described when the gunshots were fired during the walk.

Confidence: Scored from "relaxed, tail-wagging, confident posture, unaffected by the gunshots" (5), to "flees or backs away" (1). Described when the gunshots were fired during the walk.

Reaction during play: Scored from "no reaction" (5), to "flight tendencies, signs of fear" (1). Described when the gunshots were fired during play.

Confidence during play: Scored from "relaxed, tail-wagging, confident posture, unaffected by the gunshots" (5), to "flees or backs away" (1). Described when the gunshots were fired during play.

2.5. Calculation of test scores

2.5.1. Subtest scores

Several of the behavioral variables contained missing values, either because the test was stopped, or because the dog's response did not match any of the score descriptions. Similarly, three variables were scored slightly differently (i.e. scored 1, 3, or 5 instead of 1-5). To control for missing values and scoring differences, I calculated mean scores for the subtests by adding together the scores from the corresponding behavioral variables and divided by the number of scored variables (range: 1-5). The time in search variables in the subtests Search outdoors and Search indoors were measured on a different scale (i.e. in seconds), and were not included in the subtest scores.

The possibility of using subtest scores instead of behavioral variables as predictive variables depends on the assumption that each subtest measures a behavioral response in one single situation, and that subtest score reflects the corresponding variable scores. I investigated this assumption by construction of a Spearman correlation matrix between the subtests, and by comparing descriptive and inferential statistics of age differences of the subset scores and the corresponding behavioral variables. I found a low redundancy across

subtests (Appendix 1, Figure A1) and subtest scores that reflected the behavioral variables (Appendix 2, Table A2).

2.5.2. Overall score and Selected variables score

In addition to the subtest mean scores, I calculated two further mean scores: an Overall score and a Selected variables score. A dog's Overall score was calculated as an average score for the scores for the 61 discrete behavioral variables (excluding time in search). The Selected variables score was calculated as the mean score for the variables measured in the subtests that were identified by a logistic regression model to be important for predicting the test outcome at 12 months using 6-month behavior scores (section 2.6.3). The purpose of the Overall score was to give an indication of the dog's overall performance, while the Selected variables score was used to reduce possible noise caused by less predictive variables when trying to predict future behavior. The Overall score and the Selected variable score had the same range as the average subtest scores (i.e. 1-5).

2.5.3. Boldness score

Initially, I conducted my own factor analysis on the behavior scores for each measured discrete test variable (time in search excluded) using a scree plot to determine the number of extracted factors (Appendix 3, Figure A3; Table A3). This analysis identified factors corresponding to several of the personality traits observed to be related to the shyness-boldness axis (Svartberg & Forkman, 2002). However, I deemed my sample size of 62 dogs too small to be used as a foundation for further analyses. Instead, I chose to base the calculations of the factor scores on the findings of Svartberg and Forkman (2002), using only the behavior variables from the current study that were tested and measured similarly to the variables they included in their factor scores (Table A3). Due to differences between the tests, their Chase-proneness could not be evaluated from the current data. Therefore, I calculated a Boldness score for each dog was calculated based on the scores for the remaining three factors; Sociability, Playfulness, and Curiosity/Fearlessness.

Sociability was calculated using four variables from the subtest Social contact (contact with strangers, contact with tester, following, and handling). The Playfulness score was based on two variables from the subtest Playfight (intensity and drive). Curiosity/Fearlessness was calculated from three variables from the two subtests Sudden appearance and Metallic noise (startle response, exploration, and avoidance), and one variable from the subtest Ghost (exploration). To ensure that each factor score had equal weight on the Boldness score

(Svartberg, 2002; Svartberg & Forkman, 2002), I calculated the factor scores as mean values, and summed these together to create the Boldness score (range: 3-15).

Previous studies have presented their Boldness score as a standardized value (Starling et al., 2013a; Starling et al., 2013b; Svartberg, 2002), presumably because they experienced both positive and negative loaded variables on their factors (although none of the studies explicitly stated why). In this study, all the variables used to calculate factor scores had a positive loading in my initial factor analysis (i.e. a high score representing a desired behavioral response), which made summing the mean scores more intuitive given that a high Boldness score is more desirable in working dogs (Svartberg, 2002).

2.6. Statistical analyses

I assessed the assumption of normality for each average score (i.e. subtest scores, Overall score, Selected variables score, and Boldness score) using the Shapiro-Wilk test and plotting histograms. The assumption was not met, and all the statistical analyses were conducted using non-parametric tests in R version 3.5.1 (R Core Team, 2018). Most of the data preparation was done using the package tidyverse (Wickham, 2017), and the figures were made with the ggplot2 package (Wickham, 2016). Statistical differences were considered significant at $p < 0.05$.

2.6.1. Predictive validity

A chi-squared test was used to investigate any possible associations between the test outcomes (pass vs fail) at 6 and 12 months of age. However, since the chi-squared test does not test if the association is caused by chance (Whitlock & Schluter, 2015), I also ran Cohen's kappa using the 'kappa2' function in the irr package (Gamer et al., 2019). Cohen's kappa gives the percentage agreement caused by chance (Lehner, 1996), and is the recommended method for agreement assessment of categorical data (Patronek et al., 2019). The kappa coefficient (κ) is calculated by $\frac{(P_o - P_c)}{(1 - P_c)}$, where P_o is the observed proportion of agreements, and P_c is the proportion of agreements expected to occur by chance alone (Fleiss et al., 2003).

2.6.2. Test-retest reliability

Wilcoxon matched paired-tests were used to investigate if the subtest scores differed within the dogs at 6 and 12 months of age. The relationship between subtest scores and dog age when tested (days) was illustrated in scatter plots, and evaluated using Spearman rank

correlation. The level of agreement in subtest scores between the two test ages was evaluated by the Bland-Altman method, a method designed to illustrate the agreement between two paired quantitative measurements (e.g. two paired tests or methods) (Giavarina, 2015). The Bland-Altman plot is a scatter plot with the difference between two tests (T1–T2) on the y-axis plotted against the mean of the two tests $\left(\frac{(T1-T2)}{2}\right)$ on the x-axis. The bias (i.e. the consistent proclivity for the tests to be different from each other) is estimated by the estimated mean difference ($\bar{d}$), and describes the lack of agreement between the tests (Bland & Altman, 1999). The mean difference and standard deviation of the differences (s) are used to calculate limits of agreements ($\bar{d} \pm 1.96s$), which provides an 95% confidence interval for the range where most of the differences lies. The Bland-Altman plots were created using the BlandAltmentLeh package (Lehnert, 2015).

2.6.3. Predicting test outcome

I used binary logistic regression models to identify subtests associated with the dogs' test outcome at both test ages. The basic equation for a logistic regression model can be written as: $y_i \sim Binary(n_i, P_i)$

$$logit(P_i) = \log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_n X_i$$

Where the response variable (y_i) represents the test outcome (pass vs fail) for the ith dog, and P_i is the probability of a dog passing the test (y_i = 1). $\beta_n X_i$ gives the regression coefficient for the predictor variables (X_i), and the model intercept (β_0) represents the response when the predictor variable is zero (X_i = 0).

I ran three separate models. Models 1 and 2 were used to predict the test outcome at 6 and 12 months of age, respectively, while Model 3 was used to predict the test outcome at 12 months based on the mean subtest scores at 6 months. The full models had the test outcome as response variable, and subtest scores, country (Norway or Sweden), age (days) and time in search (mean time of both search subtests) as predictor variables. Model selection revealed no effect of country, age or time in search, and only subtest scores were included in the fitted models.

(1) Model 1: Test outcome at 6 months ~ Mean subtest scores at 6 months

(2) Model 2: Test outcome at 12 months ~ Mean subtest scores at 12 months

(3) Model 3: Test outcome at 12 months ~ Mean subtest scores at 6 months

Model selection was conducted using the 'stepwise' function from the StepReg package (Li et al., 2019), with significance level as the selection criterion, and entry significance level (ESL) and stay significance level (SSL) set as 0.15. The Akaike Information Criterion (AIC) was used as the information creation to identify the best fitted models. Model predictive ability was evaluated by constructing confusion matrices of predicted and actual test outcome using the 'confusionMatrix' function from the package caret (Kuhn, 2008). The predicted values were obtained by running the data through the model again using the 'predict' function in R. I set 'passed' as the positive test outcome, and calculated the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each model.

The accuracy represents the percentage of correctly predicted test outcomes (pass vs fail) and is calculated by dividing the number of correct predictions by the total number of predictions (i.e. the number of correctly predicted dogs divided by the total number of dogs).

Sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate) present the model's ability to correctly predict if a dog passed or failed, respectively (Parikh et al., 2008). Sensitivity and specificity are calculated by dividing the number of correctly predicted outcomes by the number of actual outcomes; sensitivity = $\frac{TP}{(TP+FN)}$ and specificity = $\frac{TN}{(TN+FP)}$ where the true positives (TP) and true negatives (TN) are the number of dogs that were correctly predicted as passed and failed, respectively. Similarly, the false positive (FP) and false negative (FN) represent the number of dogs wrongly predicted as passed and failed, respectively. PPV and NPV account for the prevalence of the test outcomes in the population, and represents the probability that a dog predicted as passed or failed, truly passed (PPV) or failed (NPV) (LaMorte, 2016). PPV is calculated by $\frac{Se*P}{((Se*P)+((1-Se)*(1-P)))}$ and NPV by $\frac{((Sp*(1-P))}{(((1-Se)*P)+((Sp)*(1-P)))}$ where Se is the sensitivity, Sp is the specificity, and P is the prevalence given by $\frac{(TP+FN)}{(TP+FP+TN+FN)}$.

2.6.4. Predicting future improvement

To see if it was possible to predict if a dog that failed the test at 6 months would improve enough to pass at 12 months of age, I separated the dogs into groups based on the chi-squared test: A) dogs that passed at both test ages, B) dogs that failed at 6 months, but passed at 12 months of age, and C) dogs that failed at both test ages. Dogs that passed at 6 months, but failed at 12 months, were too few to be included in further analyses (n = 3). I ran three separate generalized linear models with a one-way ANOVA design, using the groups as a

single categorical predictor variable, and Overall score, Selected variables score, and Boldness score as the response variable, respectively.

(1) Overall score ~ Groups

(2) Selected variables score ~ Groups

(3) Boldness score ~ Groups

Post hoc pairwise comparisons were conducted on differences in estimated marginal means (i.e. least square means), with Tukey p-adjustment for multiple comparisons. This was done with the ‘emmeans’ function in the package emmeans (Lenth, 2019).

3. Results

3.1. Predictive validity

I found a significant positive association between the test outcome at 6 months and the test outcome at 12 months of age ($\chi^2 = 14.78$, $p < 0.001$). Dividing the dogs into groups based on the chi-square test results, 74.2% of the dogs received the same test outcome at both test ages (Table 2). Cohen's Kappa gave a percentage agreement of 49.5 %.

Table 2. Description of four groups of dogs based on their test outcome at each test age.

Group	Test outcome at 6 and 12 months of age	N	%
A	Dogs that passed at both test ages	21	33.9
B	Dogs that failed at 6 months, but passed at 12 months	13	21.0
C	Dogs that failed at both test ages	25	40.3
D[1]	Dogs that passed at 6 months, but failed at 12 months	3	4.8

[1]Group D was excluded from further analyses due to the low sample size.

3.2. Test-retest reliability

The dogs' Overall score (the average behavior score over all subtests) differed significantly between the two test ages (mean ± SD: 6 months, 3.83 ± 0.39; 12 months, 4.01± 0.55; V = 394.5, p <0.001). The Overall score did not vary significantly across days within each test age period (Figure 1). The Bland-Altman plot revealed poor agreement between the two test ages (Figure 2). The line of zero difference does not lie within the confidence interval of the mean difference, meaning that there is a systematic difference (i.e. significant bias) between the two test ages, with higher scores at 12 than 6 months.

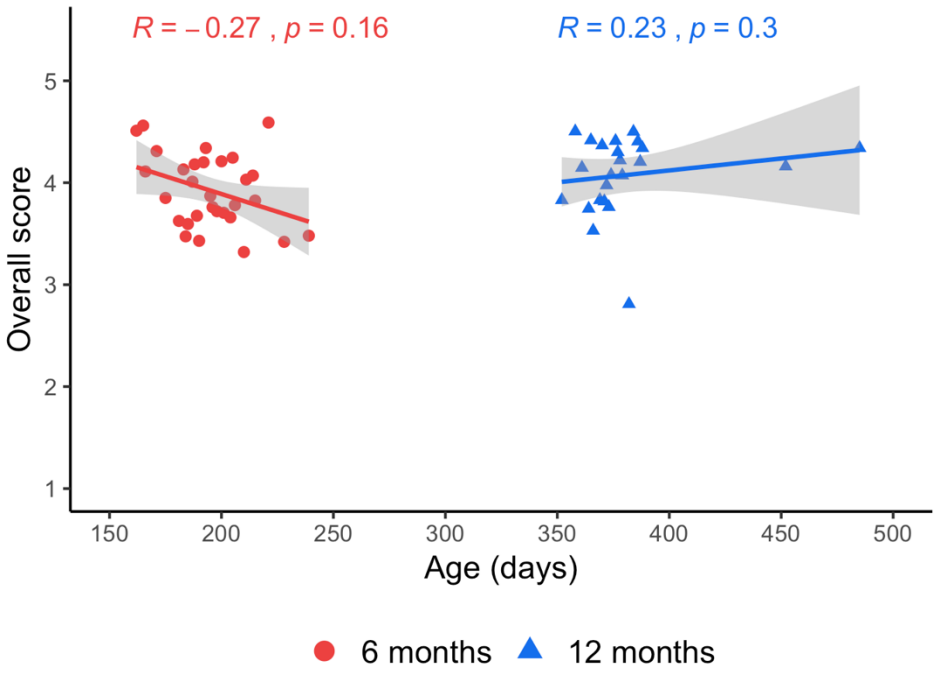

Figure 1. Association between test day and the Overall score for 62 dogs tested at approximately 6 (red dots) and 12 (blue triangles) months of age, with Spearman correlation coefficient (R), *p*-value, and regression line with 95% confidence interval (grey shading).

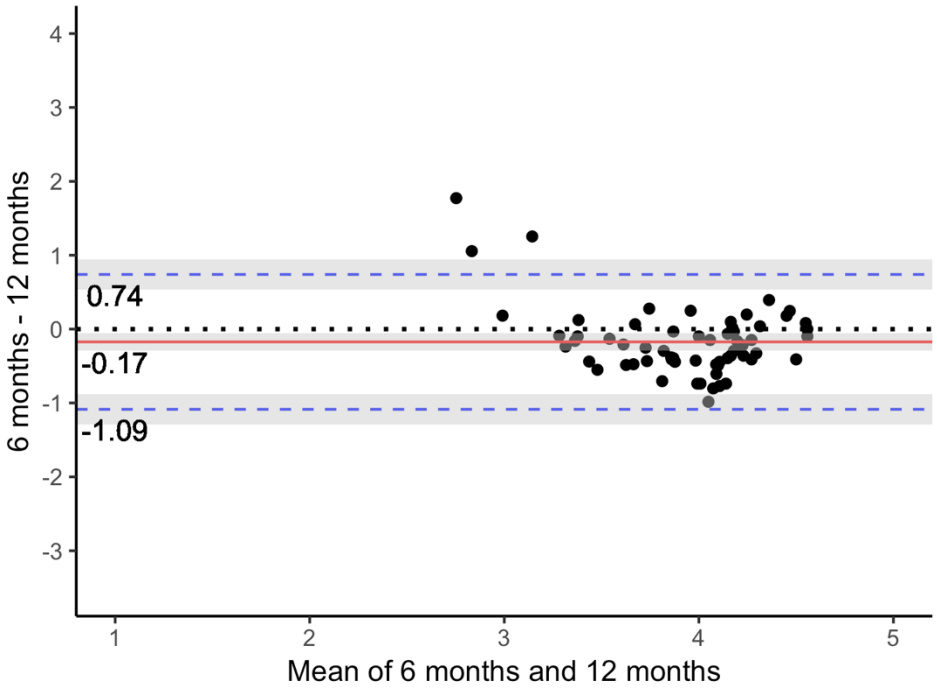

Figure 2. Bland-Altman plot of the Overall score. The y-axis shows the score difference between the two test ages. The lines represent the mean difference (red solid line), the limits of agreement (blue dashed lines), and the point of zero difference (dotted line). Light grey areas present the 95% confidence interval for the mean difference and agreement limits.

A closer examination of the individual subtests results revealed a significant difference in subtest mean scores between the two test ages in 7 of the 14 subtests (Table 3), with the dogs having a lower average score at 6 months than at 12 months of age (i.e. behavior was more desirable for police patrol work at the older test age). For most subtests, there was no significant association between test day and subtest mean score at 6 and 12 months (Figures 3 and 4). However, Social contact and Environment substrate had subtest scores that declined with test day at 6 months (Figure 3a; Figure 4e), while the Metallic noise subtest score increased with test days at 12 months (Figure 4b).

Table 3. Difference in average subtest scores between the dogs (N = 62) at 6 months and 12 months of age. Means, standard deviations, and *p*-values are given for each of the 14 average subtest scores. V represents the test statistic for Wilcoxon matched paired-test, and *p*-values < 0.05 are in bold

Subtest	6 months	12 months	Statistics	
	Mean ± SD	Mean ± SD	V	p-value
Social contact	4.15 ± 0.64	4.34 ± 0.59	490.5	**0.047**
Playfight	3.81 ± 0.50	3.95 ± 0.60	433.5	**0.032**
Retrieval	3.94 ± 0.52	4.26 ± 0.52	230.0	**0.001**
Search outdoors[1]	3.79 ± 1.10	4.19 ± 0.82	279.0	**0.003**
Sudden noise	3.94 ± 0.79	4.28 ± 0.72	272.0	**0.001**
Hunting drive	3.35 ± 0.92	4.00 ± 0.95	270.0	**<0.001**
Sudden appearance	3.45 ± 0.67	3.47 ± 0.71	663.5	0.495
Metallic noise	3.97 ± 0.70	3.99 ± 0.78	710.0	0.851
Sled	3.31 ± 0.64	3.35 ± 0.77	819.0	0.620
Ghost	3.42 ± 0.71	3.54 ± 1.02	595.5	0.145
Environment substrate	4.18 ± 0.50	4.19 ± 1.96	441.5	0.059
Dark environment	4.39 ± 0.83	4.44 ± 1.03	301.5	0.319
Search indoors[1]	3.64 ± 1.08	4.08 ± 1.16	355.0	**0.001**
Gunshot	4.40 ± 0.48	4.30 ± 0.94	711.5	0.789

[1]Time in search was measured in seconds, and was not included when calculating the subtest score.

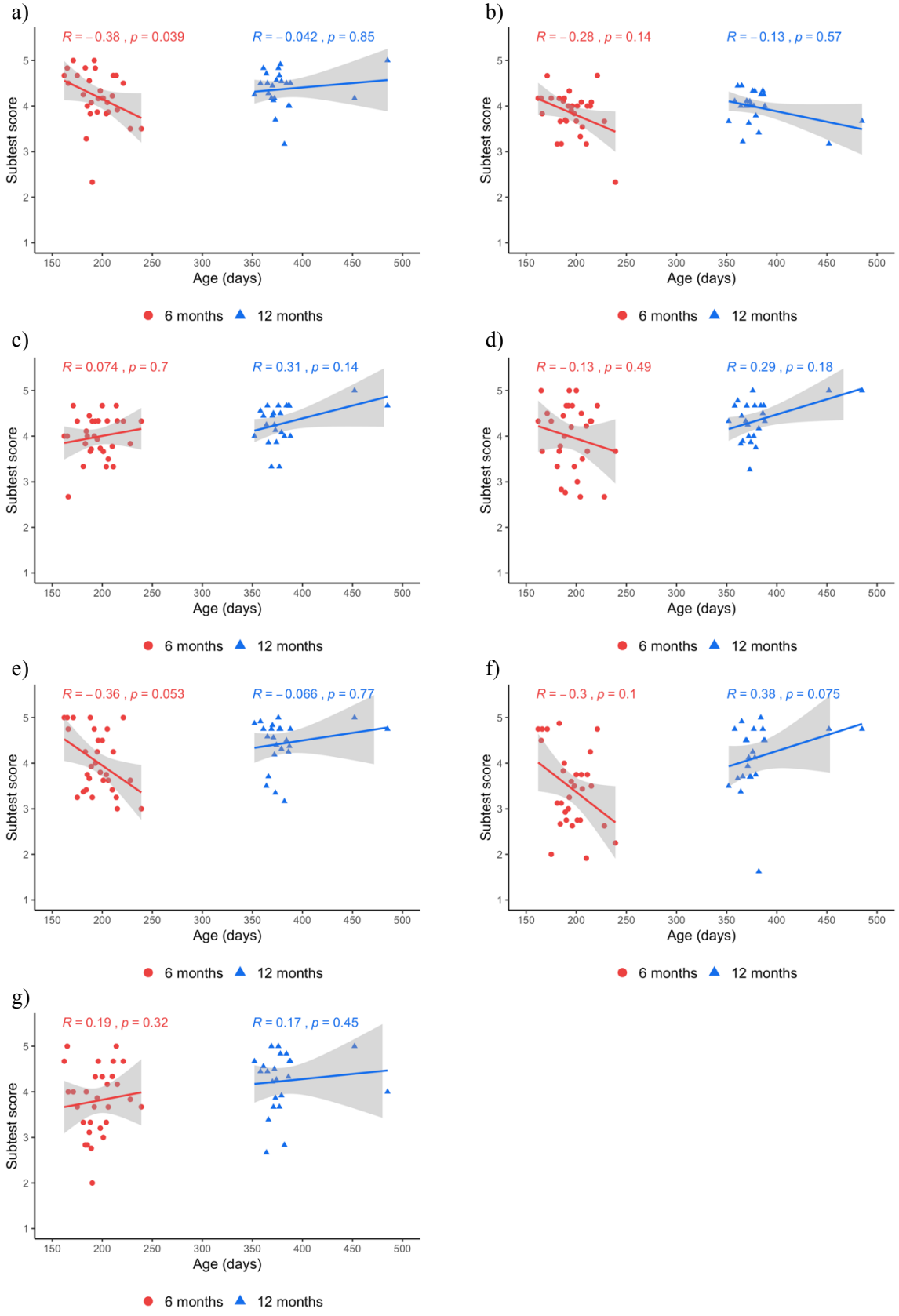

Figure 3. Association between test day and subtest score in the seven subtests with a significant age difference: a) Social contact, b) Playfight, c) Retrieval, d) Search outdoors, e) Sudden noise, f) Hunting drive, and g) Search indoors, for 62 dogs tested at approximately 6 (red dots) and 12 (blue triangles) months of age, with Spearman correlation coefficient (R), *p*-value, and regression line with 95% confidence interval (grey shading).

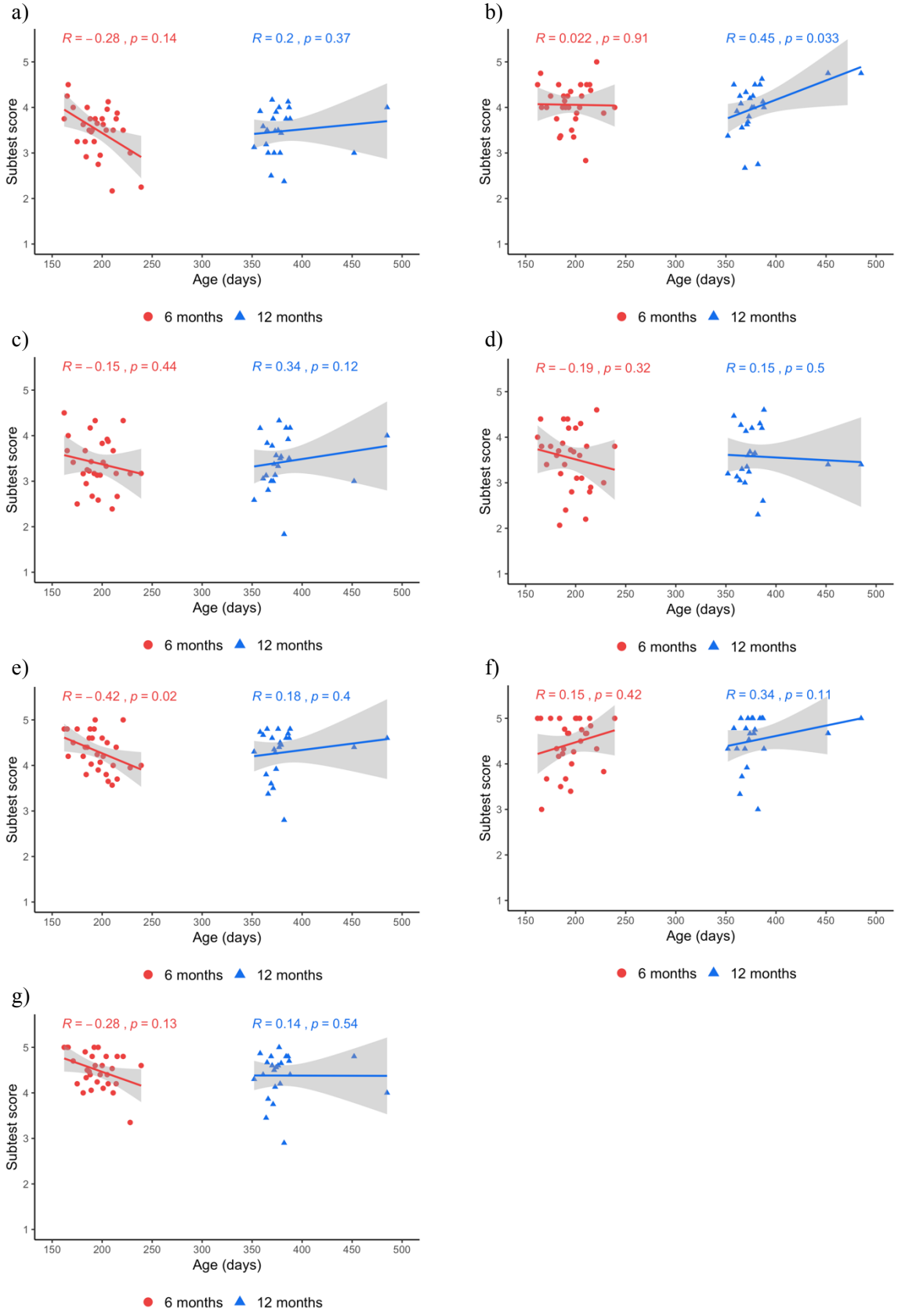

Figure 4. Association between test day and subtest score in the seven subtests without no significant age difference: a) Sudden appearance, b) Metallic noise, c) Sled, d) Ghost, e) Environment substrate, f) Dark environment, and g) Gunshot, for 62 dogs tested at approximately 6 (red dots) and 12 (blue triangles) months of age, with Spearman correlation coefficient (R), *p*-value, and regression line with 95% confidence interval (grey shading).

The seven subtests with a significant age difference (Table 3; Figure 3) also showed a significant bias between the two test ages (Figure 5). The seven subtests with no significant difference (Table 3; Figure 4) did not show this systematic difference (Figure 6). Overall, most subtest showed more points below the mean difference line than above, meaning that the dogs scored higher at 12 months than at 6 months of age. Subtests with a systematic difference had higher mean difference (i.e. bias), with Hunting drive having the highest bias (Figure 5f). Only two subtests – Playfight and Sudden noise – had all their points within the agreement limits' confidence interval, though both of these showed a significant bias between 6 and 12 months (Figure 5b; Figure 5e). Sudden appearance appears to have the best agreement level of all the subtests, with no significant bias, relatively small limits of agreements and confidence intervals, and only one point lying outside the agreement limits' confidence interval (Figure 6a).

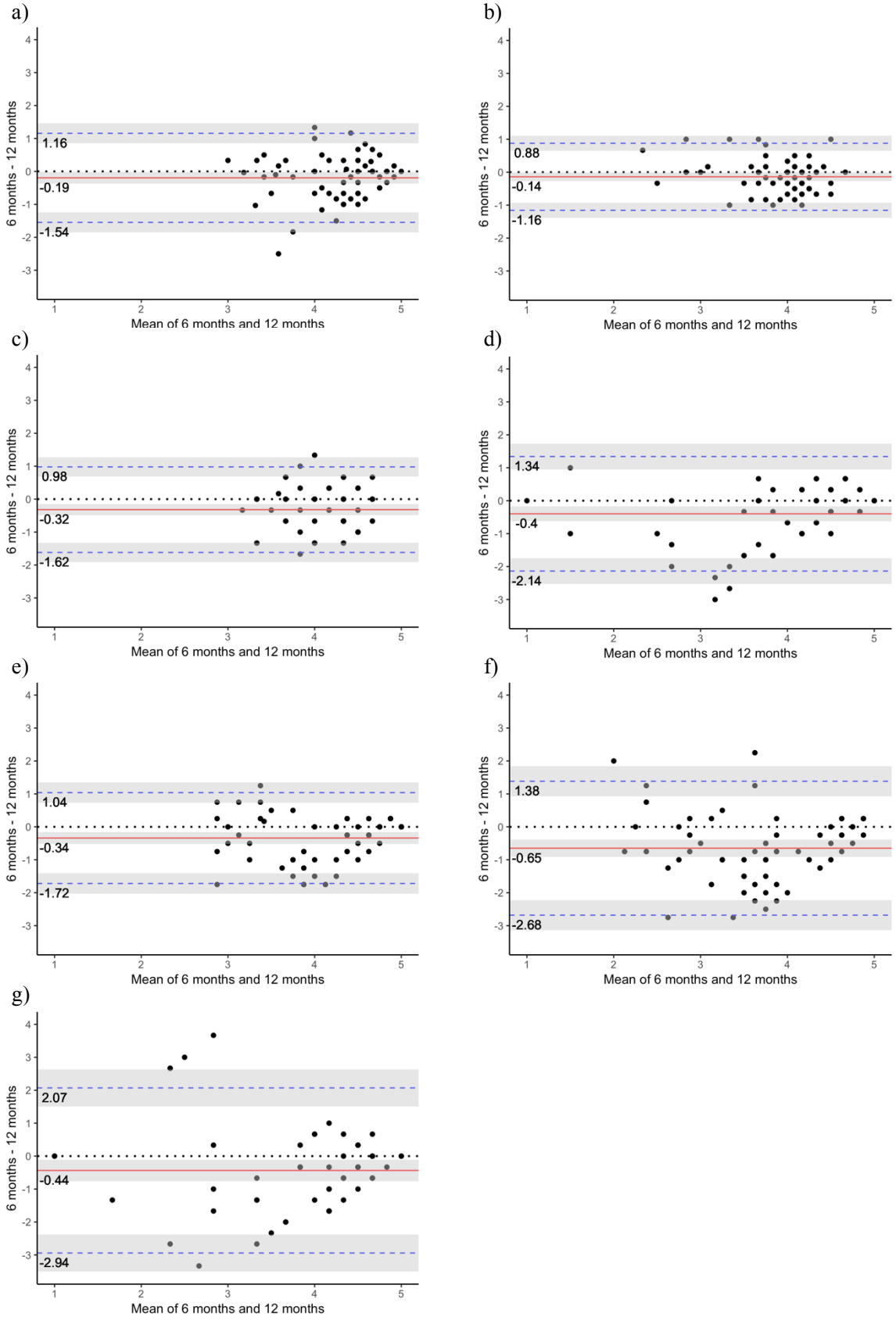

Figure 5. Bland-Altman plots of the 7 subtests with significant bias between the dogs (N = 62) at 6 and 12 months: a) Social contact, b) Playfight, c) Retrieval, d) Search outdoors, e) Sudden noise, f) Hunting drive, and g) Search indoors. The y-axis shows the score difference between the two test ages. The lines represent the mean difference (red solid line), the limits of agreement (blue dashed lines), and the point of zero difference (dotted line). Light grey areas present the 95% confidence interval for the mean difference and agreement limits.

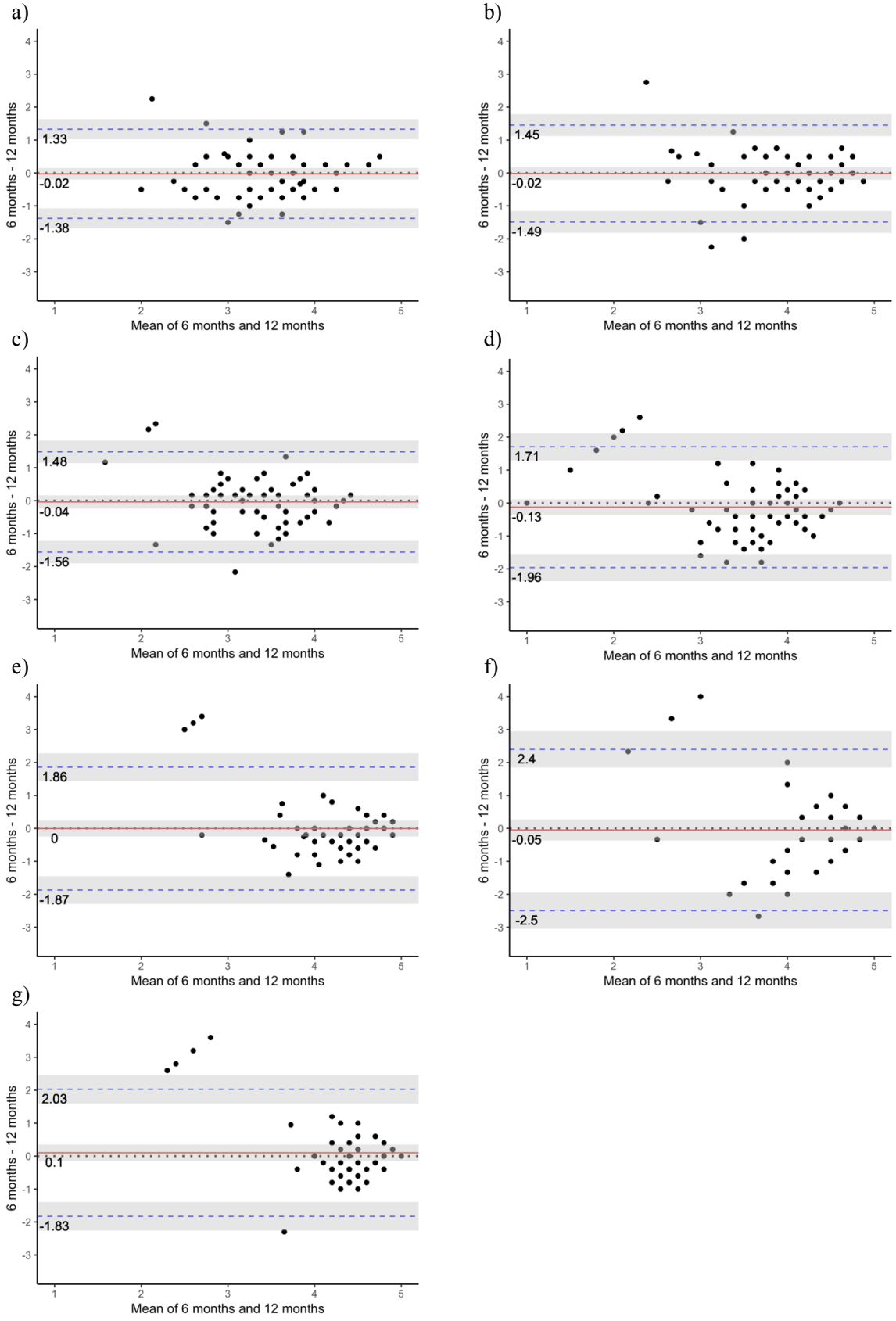

Figure 6. Bland-Altman plots of the 7 subtests with low bias between the dogs (N = 62) at 6 and 12 months: a) Sudden appearance, b) Metallic noise, c) Sled, d) Ghost, e) Environment substrate, f) Dark environment, and g) Gunshot. The y-axis shows the score difference between the two test ages. The lines represent the mean difference (red solid line), the limits of agreement (blue dashed lines), and the point of zero difference (dotted line). Light grey areas present the 95% confidence interval for the mean difference and agreement limits.

3.3. Predicting test outcome

The logistic regression models predicting test outcome at 6 and 12 months (Model 1 and 2, respectively), each revealed four subtests associated with test outcome, while the model predicting test outcome at 12 months based on the dogs' performance at 6 months of age (Model 3) had three subtests associated with test outcome (Table 4). All the subtests associated with test outcome were significant in Model 1 and 3, while three of the four associated subtests were significant in Model 2. Sudden appearance was significantly associated with the test outcome in all three models. The assessment of the models' ability to correctly predict the dogs' test outcome showed that Model 2 had the highest predictive ability, while Model 3 had the lowest (Table 5).

Table 4. Results of the fitted binary logistic regression models used to examine average subtest scores associated with test outcome (pass vs fail) at 6 and 12 months. Model 1 and 2 predict test outcome at 6 months and 12 months of age, respectively, using subtest scores from the corresponding age. Model 3 uses 6-month scores to predict test outcome at 12 months of age. Z represents the model test-statistic, and *p*-values > 0.05 are in bold.

Predictor variable	Model 1		Model 2		Model 3	
	z	p-value	z	p-value	z	p-value
Social contact	-	-	-	-	2.004	**0.045**
Playfight	-2.588	**0.010**	-	-	-	-
Sudden appearance	2.758	**0.006**	2.475	**0.013**	3.302	**0.001**
Search outdoors	2.532	**0.011**	-	-	-	-
Ghost	-	-	2.157	**0.031**	-	-
Environment substrate	2.451	**0.014**	-	-	-	-
Dark environment	-	-	-1.796	0.072	-	-
Search indoors	-	-	2.992	**0.003**	2.676	**0.007**

Table 5. Statistics used to evaluate the models' ability to correctly predict the dogs' test outcome, including model accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Model 1	85.5	79.2	89.5	82.6	87.7
Model 2	91.9	94.1	89.3	91.4	92.6
Model 3	79.0	82.4	75.0	80.0	77.8

3.4. Predicting future improvement

Pairwise comparisons of differences in estimated marginal means (i.e. least square means) revealed that, at 6 months, the Overall score, the Selected variables score, and the Boldness score differed significantly between the three groups obtained from the chi-squared test (Table 2), with Group A having the highest score and Group C the lowest (Figure 7).

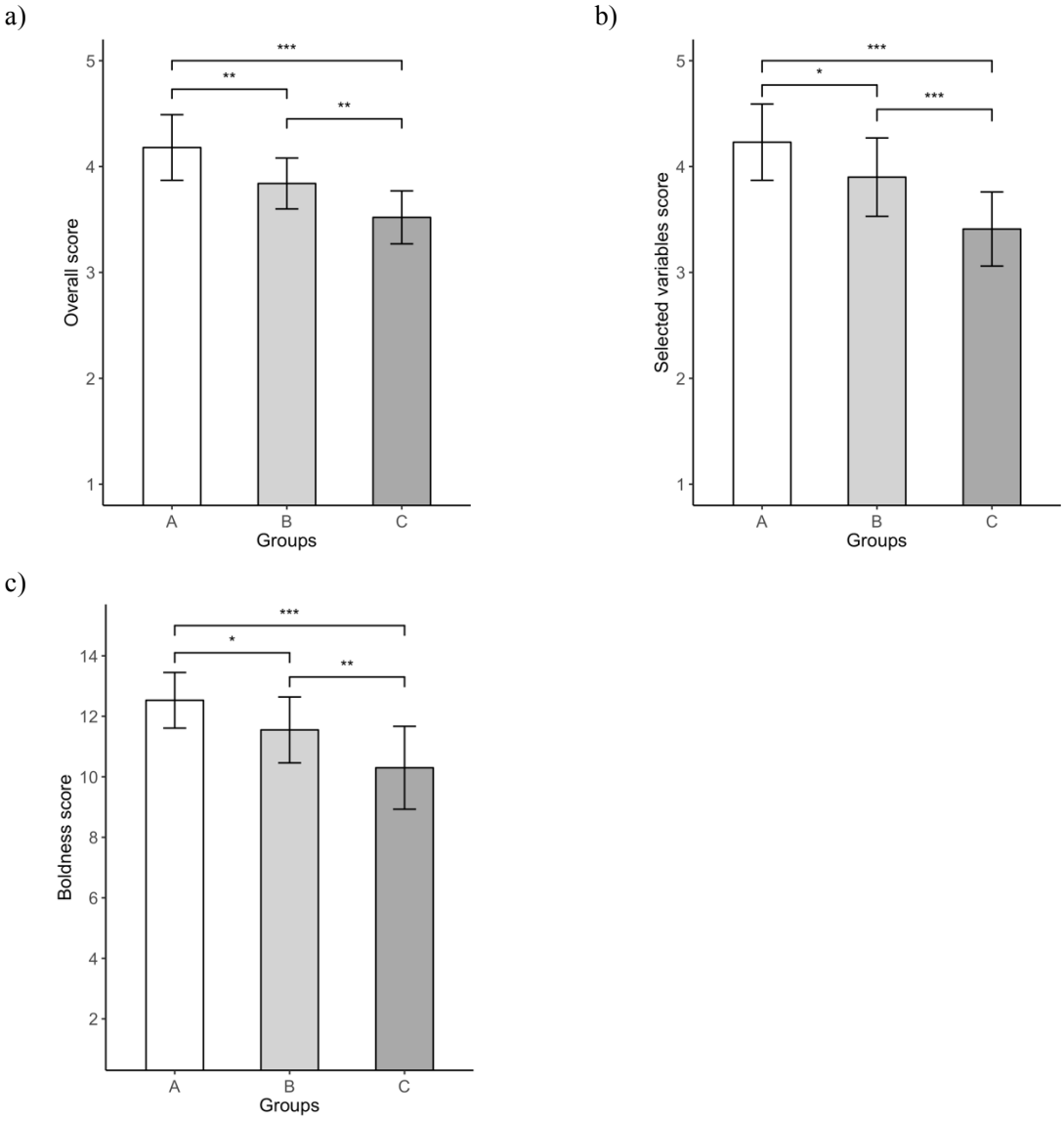

Figure 7. Differences at 6 months between the dogs that passed at both test ages (Group A, n = 21), the dogs that failed at 6 months, but passed at 12 months (Group B, n = 13), and the dogs that failed at both test ages (Group C, n = 25). The x-axis shows the average a) Overall score, b) Selected variables score, and c) Boldness score, with significance level above, and error bars representing the standard deviation.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Wilcoxon matched paired-tests showed that Group B was the only group with scores that improved (i.e. increased) significantly from 6 months to 12 months of age (Overall score: V = 0, $p < 0.001$; Selected variables score: V = 3.5, $p = 0.004$; Boldness score: V = 9, $p = 0.012$) (Figure 8). At 12 months, Group C scores were significantly lower than both Group A scores (Overall score: z = -5.741, $p < 0.001$; Selected variables score: z = -5.974, $p < 0.001$; Boldness score: z = -5.926, $p < 0.001$) and Group B scores (Overall score: z = -5.417, $p < 0.001$; Selected variables score: z = -5.240, $p < 0.001$; Boldness score: z = -4.446, $p < 0.001$).

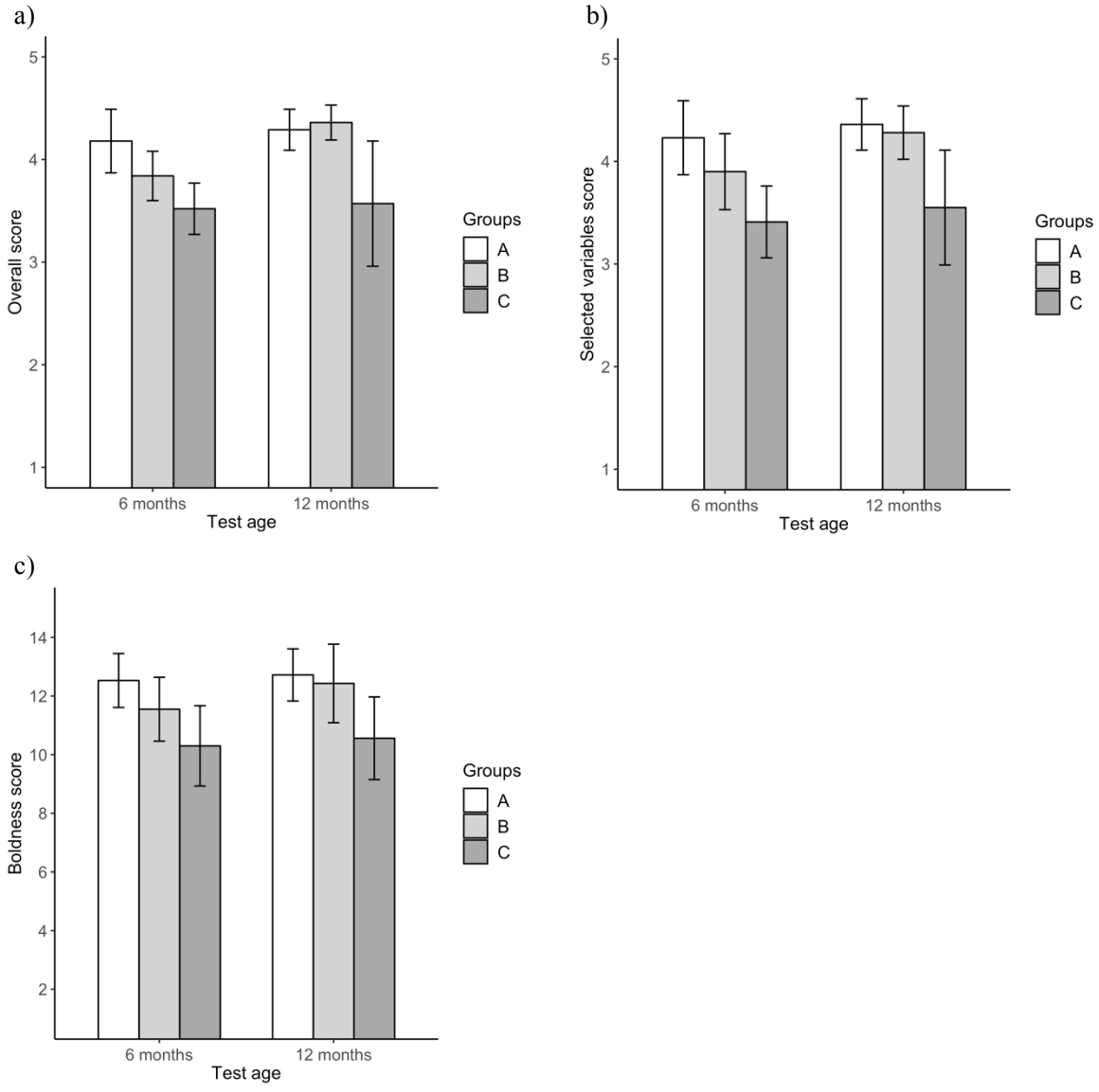

Figure 8. The average a) Overall score, b) Selected variables score, and c) Boldness score at 6 and 12 months for the dogs that passed at both test ages (Group A, n = 21), the dogs that failed at 6 months, but passed at 12 months (Group B, n = 13), and the dogs that failed at both test ages (Group C, n = 25). Error bars represent the standard deviation.

4. Discussion

4.1. Overview

The aim of this study was to investigate the possibility of assessing police dog qualification at 6 months instead of at 12 months of age. When comparing the test outcome (pass vs fail) from the two test ages, I found a strong association with a high percentage (74.2%) of dogs receiving the same test outcome at 6 months and 12 months of age. Furthermore, Cohen's Kappa gave a percentage agreement of 49.5%, suggesting that test outcome at 6 months had a moderate predictive validity of test outcome at 12 months of age. Assessment of test-retest reliability revealed evidence of temporal consistency in 7 of the 14 subtests, though scatter plots showed that individual variation was high in all subtests at both test ages. Binary logistic regression models identified four and three subtests associated with test outcome at 6 months 12 months of age, respectively. Additionally, three subtests were significantly associated with test outcome at 12 months when using behavior scores from 6 months of age. Only one subtest – Sudden appearance – was significantly associated with test outcome in all three models. Lastly, I found that, at 6 months, dogs that failed the qualification test at 6 months, but passed at 12 months had a significantly higher Overall score, Selected variables score, and Boldness score compared to the dogs that failed at both test ages. Below, I discuss these results in more detail.

4.2 Predictive validity

Predictive validity describes how well a measurement predicts future performance (e.g. how well the result of a behavioral test reflects the outcome at a later test) (Lin & Yao, 2014; Patronek et al., 2019). I found that test outcome had an agreement of 49.5 %, which is considered to represent a moderate level of agreement (Altman, 1991; Fleiss et al., 2003). The chi-square test showed that test outcome at 12 months was not independent from the 6-month outcome, with 74.2% of the dogs receiving the same test outcome at both test ages. This percentage was higher than the agreement percentage, which is expected when correcting for change (Patronek et al., 2019). My findings are somewhat similar to the findings of Wilsson and Sundgren (1997), who reported that approximately 50% of the German shepherds and Labrador retrievers (450-600 days) that passed the qualification test successfully completed training. In contrast, Svobodová et al. (2008) reported that 71.8% of the German shepherds tested as puppies (7 weeks) passed the qualification test as adults (age not specified).

However, 78.7% of the dogs in the initial sample size did not return for the qualification test for various reasons, including undesirable behavior (Svobodová et al., 2008), suggesting that the predictive validity was in realty much lower.

4.3. Test-retest reliability (temporal consistency)

4.3.1. Association between test day and subtest score

The scatter plots revealed individual variation at both test ages, and despite the regression lines suggesting strong associations between test age (days) and average subtest scores, only three subtests had average scores associated with the test age: Social contact and Environment substrate at 6 months (Figure 3a; Figure 4e), and Metallic noise at 12 months (Figure 4b). Overall, the correlation coefficients were low to moderate, and should be viewed with caution.

4.3.2. Assessment of temporal consistency

There is some confusion between agreement and correlation (Kalra, 2017; Patronek et al., 2019). Test-retest reliability is often defined as the correlation between two measures (e.g. Svartberg, 2005), though a strong correlation does not necessarily imply a high agreement level (Giavarina, 2015; Patronek et al., 2019). However, good correlation is important between two measurements of the same variable (Kalra, 2017). It is also worth mentioning that, while comparison of means is not a measure of agreement (Whitlock & Schluter, 2015), the seven subtests with average scores that differed between the two test ages, were the same subtests with poor agreement level (i.e. significant bias) (Table 3; Figure 5). This suggests that it is possible to obtain similar results despite implementations of different methods, which in turn, suggests that comparisons of findings despite use of different methods may be useful.

It is important to note that, while an Bland-Altman plot generates agreements limits, it does not provide information about whether these limits are acceptable (Kalra, 2017). The limits should ideally be defined before (*a priori*) the conduct of statistical analyses, and according to biological or other relevant factors (Giavarina, 2015).

The assessment of test-retest reliability (temporal consistency) is also affected by the age at first measurement and the interval length between measurements (Fratkin et al., 2013). To my knowledge, there is only one other study assessing dogs' temporal consistency by repeated behavioral testing in a similar age interval (5-8 months) (Harvey et al., 2016b). While personality consistency in dogs has received interest over the last few decades, little

has been done on working dogs (Fratkin et al., 2013), where most studies focus on identifying personality associated with successful training completion (e.g. Svobodová et al., 2008; Wilsoon & Sundgren, 1997; Wilsson & Sinn, 2012). However, since temporal consistency is necessary when predicting future behavior (Svartberg et al., 2005), such studies provide some information about personality consistency within their respective age intervals.

4.3.3. Subtests without temporal consistency

Seven subtests – Social contact, Playfight, Retrieval, Search outdoors, Sudden noise, Hunting drive, and Search indoors – did not show temporal consistency in the interval 6-12 months of age (i.e. the average subtest scores showed significant bias between the two test ages). The average subtest scores for these subtests were higher at 12 months. In other words, the dogs showed more desirable police patrol dog behavior at the older age, which corresponds with my prediction (Table 1).

My finding of no temporal consistency in the subtest Social contact (i.e. sociability), is somewhat consistent with previous findings in guide dogs (mostly crossbreeds of Labrador retrievers and Golden retrievers) in the interval 5-8 months of age (Harvey et al., 2016b). Harvey and her colleagues (2016b) reported consistency in low body posture when greeting a stranger, but they found no evidence of consistency when assessing compliance during a body check subtest. Their 'low body posture' variable is similar to my 'contact with strangers' and 'contact with tester' variables, while their 'body check' corresponds with the variable 'handling' in the current study. The mixed results on temporal consistency regarding sociability in juvenile dogs (5-12 months) could be affected by personality differences reported between German shepherds and Labrador retrievers (Wilsson & Sundgren, 1997). On the other hand, fear towards strangers (i.e. reluctance to approach) was significantly higher in dogs reported to have been frightened by an unfamiliar or familiar person at an earlier age (Serpell & Duffy, 2016), and sociability might, therefore, vary in consistency due to individual experiences.

Higher resilience during tug-of-war (Playfight) and higher desire to chase and fetch a ball (Retrieval) in 7 weeks old German shepherds increased the probability of passing police dog qualification (Svobodová et al., 2008). Similarly, the willingness to engage in tug-of-war, as well as the persistency during the game, were higher in adult German shepherds (range 15-20 months) that successfully completed 8-10 months of training as police dogs (Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012). This suggests that behavior during tug-of-war is consistent in dogs, which contradicts with my findings. This difference could be caused by

comparison of subtest scores instead of individual behavioral variables. The significant age difference in the subtest Playfight appears to be largely influenced by the difference in one behavioral variable (confidence), while the other variables measuring resilience and intensity were not significantly different between the two test ages (Appendix 2, Table A2). It is possible that the 'confidence' variable affected the subtest score enough that it did not reach an acceptable agreement level (Figure 5b), which suggests that not all subtest scores are as representative of the corresponding variables as I had assumed. On the other hand, Svobodová et al. (2008) and Wilsson and Sinn (2012) also used aggregated scores, obtained by factor analysis (FA) and principal component analysis (PCA), respectively. The combined score of tug-of-war and ball retrieval (Svobodová et al., 2008), showed great variation between individual puppies, and while Wilsson and Sinn (2012) lacked information on individual variation, their aggregated score consisted of several, unrelated variables, making it impossible to assess the consistency of tug-of-war alone. Slabbert and Odendaal (1999) assessed retrieval behavior in German shepherds and while they found that higher performance in a retrieval test at both 8 and 12 weeks of age increased the chance of becoming a successful police dog at 18-24 months of age, the scores appeared to change from 8 to 12 weeks, indicating low consistency.

Focus and persistence during search, both Search outdoors and Search indoors, did not show temporal consistency in this study. These subtests assessed the dogs' ability to keep focused during the designated task, which could be affected by age. A questionnaire study investigating dogs' attention skills, indicated that younger dogs (10-24 months) had higher inattention scores compared to older dogs (> 2 years) (Vas et al., 2007). The lack of consistency during search could also have been affected by additional training. The dog's keeper was present during the tests, and could have increased training after watching their dog perform at 6 months of age. This would be consistent with other findings of Vas and her colleagues (2007), where the dogs' inattention scores were lower in dogs that had received systematic training, regardless of age.

Sudden noise measured the dogs' reaction toward a sudden noise and their willingness to approach and explore the source. There is evidence that German shepherd subjected to auditory stimulation during development (16-32 days) responded less to a sudden noise subtest at 7 weeks of age during police dog selection testing in the Czech Republic (Chaloupková et al., 2018). However, in the current study, the dogs' reaction to the noise was very similar at 6 months and 12 months of age (Appendix 2, Table A2). The lack of temporal consistency appears instead to be affected by the dogs' behavior when unleashed (curiosity)

and their confidence during the approach (Table A2). Previous experiences are known to affect behavior in dogs (e.g. Serpell & Duffy, 2016), and it is possible that the dogs remembered the subtest from when first tested at 6 months, resulting in a faster and more confident approach at 12 months of age.

There is evidence that the willingness to run after a moving object (Hunting drive) is consistent in adult dogs (1-2 years) across both longer and shorter intervals (Svartberg, 2005; Svartberg et al., 2005). My findings provide evidence that this is not the case in the interval 6-12 months, where Hunting drive had the highest bias of all the subtests when assessing level of agreement (Figure 5f). My findings are somewhat similar to findings of Harvey and her colleagues (2016b), who reported mixed results when testing approach or avoidance of two different fake birds pulled on a lead in front of juvenile guide dogs. In the interval 5-8 months, they found approaching or avoiding a singing robin soft toy showed temporal consistency, while the same test using a pair of decoy pigeons did not (Harvey et al., 2016b).

4.3.4. Subtests with temporal consistency

In this study, I found evidence of temporal consistency in the remaining seven subtests; Sudden appearance, Metallic noise, Sled, Ghost, Environment substrate, Dark environment, and Gunshot.

The first four subtests measured startle response and recovery when exposed to different stimuli. Defense was directly assessed in Sudden appearance, Sled, and Ghost, while the Metallic noise subtest measured flight tendencies. In a study of adult dogs (1-2 years), Svartberg (2005) found the personality trait Curiosity/Fearlessness to be consistent in longer interval (duration 1-2 years). The study used data from 697 dogs of 16 breeds, including German shepherds (Svartberg, 2005). The majority of the variables measured in the subtests Sudden appearance and Metallic noise (startle response, exploration, and avoidance) in the current study were included in the aggregated Curiosity/Fearlessness score in Svartberg's study (2005). This suggests that my findings of temporal consistency in the subtests Sudden appearance and Metallic noise might not only apply to the 6-12 interval, but also remain consistent in adulthood.

The subtest Sled was not included in Svartberg's study (2005). However, in the current study, several of the behavioral variables measured in the Sled subtest were also measured in the Sudden appearance and Metallic noise subtests. Furthermore, the average score of the subtest Sled had moderate correlation ($r_s \geq 0.50$) with the average subtest scores for Sudden Appearance and Metallic noise (Appendix 1, Figure A1). There is, therefore,

possible that the temporal consistency in the subtest Sled also might be applied extend to more than the 6-12 interval found in this study.

Ghost was the only subtest involving human approach or threat, and the consistency observed in this study is perhaps surprising considering that aggression towards people has been reported to increase in German shepherd between 6 and 12 months (Serpell & Duffy, 2016). It is possible that the dogs evaluated as guide dogs are less bold than those dogs subjected to police dog training. Wilsson and Sundgren (1997) found that German shepherds selected as guide dogs tended to react less aggressively than the German shepherds selected as police dogs. Furthermore, Serpell and Duffy (2016) evaluated both females and males, whereas only males were evaluated in the current study. In dogs, males are generally bolder than females (Kubinyi et al., 2009; Starling et al., 2013a), and tend to exhibit more aggressive behavior (Miklósi, 2015). There is also a relationship between aggression and fear (Miklósi, 2015), and threats can be an expression of fear (Gray, 1987). There is evidence that fearfulness at 3 months of age in four breeds of guide dogs (including German shepherds) is somewhat predictive of adult behavior (Goddard & Beilharz, 1984). Foyer and her colleagues (2014) found that German shepherds exhibiting fearful behavior (i.e. stranger-directed fear, dog-directed fear, and non-social fear) at 14 months of age, had a smaller chance of passing a qualification test at 17 months of age. Therefore, it is possible that dog keepers with dogs exhibiting fear at an early age (~3 months) chose not to participate in this study, due to the reduced chances of passing the qualification test.

The Environment substrate and Dark environment subtests assessed the dog's environmental sureness (i.e. ability to move around confidently in an unfamiliar and distracting environment). Reimer and her colleagues (2014) conducted a longitudinal study of pet dogs and found exploration of an unfamiliar room to be affected by age. The dogs spent a lot more time moving and exploring (sniffing various surfaces) the unfamiliar room when they were adults (1.5-2 years) compared to when they were puppies (40-50 days) (Riemer et al., 2014). My findings of temporal consistency conflict somewhat with the findings of Reimer et al. (2014), though this is probably due to differences in age and the duration of the interval between tests (Fratkin et al., 2013). The use of a standardized behavior test provides the dog keeper with knowledge of the different subtests their dog will encounter, and it is possible that the consistency observed in the current study was affected by previous experience. Reduced responsiveness in a situation (or to a stimulus) can be obtained by repeated exposure (McFarland, 2006). Acclimation to a new environment has been observed

to affect behavior in shelter dogs, where sociability towards unfamiliar people increased with time (days) since arrival (Goold & Newberry, 2017b).

Reaction to gunshots is a common feature in behavior tests in different working dog programs (e.g. Foyer et al., 2013; Slabbert & Odendaal, 1999; Svartberg, 2002). Despite this, gun sureness in working dogs (German shepherds age 12-18 months) has not been found related to boldness (Svartberg, 2002) or aggregated scores associated with passing a qualification test or completing training (Foyer et al., 2013; Wilsson & Sinn, 2012). However, gun sureness is important trait in workings dogs, and exhibiting fearful behavior when exposed to gunshots has been considered enough to disqualify German shepherds (450-600 days) from becoming police dogs (Wilsson & Sundgren, 1997). Assessment of gun sureness in German shepherds at 12 weeks of age provided little predictability of later success in police dogs (Slabbert & Odendaal, 1999). My findings in the current study suggest that gun sureness in German shepherds can be assessed at 6 months of age.

4.4. Predicting test outcome

Three binary logistic regression models were used to identify subtests with an average score associated with test outcome (pass vs fail). Models 1 and 2 assessed subtests associated with test outcome at 6 months and 12 months, respectively, using average subtest scores from the corresponding age. Model 3 used subtest scores from 6 months to investigate subtests associated with test outcome at 12 months of age and was, thus, the only model that predicted future test outcome. Contrary to my prediction of temporal consistency between 6-month subtests and 12-month test outcome, only Sudden appearance showed evidence of temporal consistency in this study.

4.4.1. Subtests associated with test outcome

In this study, only one subtest – Sudden appearance – was significantly associated with the test outcome in all three models. The tendency of a dog to defend itself or its handler, and the ability to overcome and recover from fearful and stressful situations (often referred to as ‘defense drive’ and ‘nerve stability’), were found to be higher in 450-600 day-old German shepherds that later successfully complete police dog training (Wilsson & Sundgren, 1997). A subsequent study compared the results from a standardized behavior test of 15-18 months old German shepherds to training outcome (success or rejection) of working dogs in the Swedish Armed Forces (SAF) program (Wilsson & Sinn, 2012). In contrast to Wilsson and Sundgren

(1997), Wilsson and Sinn (2012) found no association between defense drive and training completion, though they did find that dogs exhibiting less fearful behavior and showing higher nerve stability, had a higher probability of successfully completing training (Wilsson & Sinn, 2012).

In the current study, behavior during search was also significantly associated with test outcome in all three models (Search outdoors in Model 1; Search indoors in Models 2 and 3). This is somewhat surprising, considering that focus during searches was found to be a more central trait in adult detection dogs compared to patrol dogs (Goold et al., 2016). Higher intensity and persistence during search increased the odds of completing training in German shepherds in the SAF program, where working role was determined after training completion (Wilsson & Sinn, 2012).

In this study, environmental sureness was assessed in two subtests; Environment substrate and Dark environment. I found that, in Models 1 and 2, the average score for the Environment substrate subtest was significantly associated with test outcome at 6 months of age, whereas the Dark environment score was not associated with test outcome at 12 months. The subtest Environment substrate measured the dog's behavior on stairs and slippery floor. Fearlessness (both social and non-social) is a desirable trait in police dogs (Goold et al., 2016; Wilsson & Sundgren, 1997), and it has been suggested that stairs-related fear develops independently from fear related to other non-social stimuli (e.g. sound, objects, and novel situations) (Serpell & Hsu, 2001). Furthermore, the ability to move on a slippery floor is associated with the ability to adapt to novel situations in patrol dogs (Goold et al., 2016), which is a necessary trait in police and military dogs placed in active service (Slabbert & Odendaal, 1999; Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012).

Play with handlers (tug-of-war) was found to be a central personality trait among adult police patrol dogs (Goold et al., 2016), and it is surprising that the average score of the subtest Playfight was only associated with test outcome at 6 months in this study. Evidence suggests that playfulness in dogs is correlated with other variables (e.g. obedience) (Bradshaw et al., 2015). Goold et al. (2016) reported positive correlations between play and curiosity, and between curiosity and fearlessness, both of which were somewhat measured in the subtest Sudden appearance. It is possible that the average score for the subtest Playfight was not associated with the test outcome in the other models because desirable traits measured in the subtest Playfight were also measured in the subtest Sudden appearance. Similarly, the subtest Sudden appearance assessed several of the same behavioural variables as the subtest Ghost, which could explain why the average score of Ghost was only associated with test outcome in

one model, despite the desirableness of the tendency for the dog to defending itself and the handler (Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012).

Model 3 identified three subtests with average 6-month scores associated with test outcome at 12 months; Social contact, Sudden appearance, and Search indoors. There are conflicting results regarding the predictive value of sociability in working dogs. In Wilsson and Sinn's study (2012) of 15-18 months old German shepherds, sociability was included in the aggregated score reflecting the dogs' confidence, which was associated with test outcome. Specifically, higher confidence increased the odds of completing training (Wilsson & Sinn, 2012). In contrast, Wilsson and Sundgren (1997) found no association between training completion and sociability in German shepherds. Showing appropriate behavior when meeting strangers is undoubtedly a desired trait in a police dogs, and dogs exhibiting fearful behavior (i.e. stranger-directed fear, dog-directed fear, and non-social fear) have a smaller chance of passing the qualification test (Foyer et al., 2014). However, negative experiences can influence sociability (Serpell & Duffy, 2016), and the conflicting results of the predictive value of sociability assessment in working dogs could be caused by variations in the different dog populations.

The results in the current study (Table 4) raise the possibility of a shorter test procedure. However, there is evidence that many desirable traits in police patrol dogs are correlated with each other (Goold et al., 2016). Similarly, many studies predicting test outcome have used aggregated scores obtained from multivariate analyses (e.g. Foyer et al., 2014; Harvey et al., 2016b; Wilsson & Sinn, 2012). This suggests that suitability of police dogs is assessed based on behavior scores across subtests. Dogs' behavioral responses differ across contexts (Goold & Newberry, 2017a), and the use of a shorter test procedure could result in loss of important information.

4.4.2. The models' predictive ability

Predictive ability describes the accuracy of the method used to predict future performance (Patronek et al., 2019). In the current study, 6-month behavior scores correctly predicted 79.0% of the dogs' test outcome at 12 months (Table 5). This is consistent with findings in 5-month-old guide dogs (mostly crossbreeds of Labrador retrievers and Golden retrievers), where the logistic regression model had a 79.7% accuracy when predicting withdrawal or completion of training (Harvey et al., 2016b). Similarly, assessment of aggression in 6-month-old German shepherds correctly predicted failure as police dogs (i.e. specificity) for 78.1% of the dogs (Slabbert & Odendaal, 1999). Furthermore, a logistic regression model using

behavior scores from 15-18 months old German shepherds yielded an accuracy of 78.3% when predicting training completion in working dogs (Wilsson & Sinn, 2012). However, it is important to note that the predicted values were obtained by running the same data through the models a second time. This will most likely cause overfitting (Han et al., 2012), which refers to the model's inclusion of random effects. Therefore, results might be specific to that population and not generalizable to other populations of police dogs. Thus, the predictive ability listed in Table 5 represents the 'best-case scenario', and would most likely be lower if the models were used to analyze new data from other police patrol dog populations.

In this study Model 2 had a higher predictive ability than the other models (Table 5). This could be because the dogs were older and, thus, scored more consistently across subtests. Another possibility is related to the experience of the testers. In existing protocols, dogs were test at 12 months of age in Norway whereas tests were conducted at 18 months in Sweden. Evidence suggests that dog personality is more consistent after 12 months of age (Fratkin et al., 2013), and the behavior observed at 12 months of age might have, therefore, been more similar to what the Norwegian and Swedish testers were used to assessing.

4.5. Predicting future improvement

One of the challenges with testing of working and service dogs at younger ages is the possibility of losing potentially good dogs because they lack the desired personality at the earlier test age. In this study, I have found a difference between dogs that failed qualification at 6 months, but passed at 12 months (Group B) and the dogs that failed at both test ages (Group C). This supports my prediction that dogs that would pass the qualification test at 12 months already showed a detectable potential at 6 months of age.

In contrast to prediction of future test outcome, assessment of future improvement (i.e. higher behavior scores at 12 months) requires higher scores at 12 months (i.e. low temporal consistently). In the current study, significant bias between the test ages was found in the Overall score (Figure 2), and the Boldness score and Selected variables score (Appendix 4, Figure A4). Dogs that passed qualification showed more boldness and achieved higher scores across all subtests (i.e. higher Boldness score and Overall score) compared to dogs that failed, regardless of test age (Figure 8), which is consistent with previous findings (Svartberg, 2002). The Selected variables score was calculated to measure potential improvement (i.e. higher behavior scores at 12 months) without the additional noise from subtests with less predictive value (i.e. the Overall score). This was successful, and at 6 months of age, I observed a larger difference between Group B and Group C for the Selected variables score compared to the

Overall score and the Boldness score (Figure 7). Since Group B also achieved a higher Overall score and Boldness score than Group C at 6 months, the Selected variables score can be considered somewhat representative for behavior scores across the all subtests, as well as the dog's boldness. This suggests that the subtests included in the Selected variables score (i.e. Social contact, Sudden appearance, and Search indoors) can be used as a potential guide for testers to use as an indicator of future improvement.

Using statistical analyses to predict future test outcome using test scores has suggested as a method of selection (Harvey et al., 2016b). However, this would require additional training within the police. My findings provide evidence that potential police dogs will show potential at 6 months of age, even if they do not pass the qualification test. Good inter-tester reliability between testers with minimal training and experts has been reported when assessing military working dogs (Fratkin et al., 2015). This suggests that testers may not require extensive to be able to identify dogs with potential for future improvement.

4.6. Practical considerations

Personality in dogs varies between sexes and breeds (Goold & Newberry, 2017a; Hart & Hart, 2017; Serpell & Hsu, 2005). Wilsson and Sundgren (1997) used a standardized behavioral test to investigate personality differences between German shepherds (police dogs) and Labrador retrievers (guide dogs) and found, when excluding sex, a difference in 8 of the 10 characteristics they measured. When comparing test results from all dogs, regardless of working role or training success, they found that German shepherds had a higher tendency to react aggressively (sharpness) and were more likely to defend their handler or themselves (defense drive), whereas the Labrador retrievers were more cooperative and social, and showed a greater ability to overcome and recover from fearful or stressful situations (nerve stability) (Wilsson & Sundgren, 1997). Furthermore, males of both breeds showed less fearful behavior and higher defense drive, and male German shepherds also exhibited higher nerve stability compared to females (Wilsson & Sundgren, 1997). Only male German shepherds were tested in the current study, and my findings might not be generalizable to female German shepherds or other breeds.

When investigating the correlation between boldness and working dog performance, Svartberg (2002) found that German shepherds had a higher Boldness score than Belgian tervuren regardless of sex, though males of both breeds had higher boldness compared to females (Svartberg, 2002). This sex difference has also been reported in companion dogs (Kubinyi et al., 2009; Starling et al., 2013a). Additionally, studies of companion dogs have

also found that boldness varies between breeds (Starling et al., 2013b), and is affected by reproductive status, with intact dogs being bolder than neutered dogs (Bennett & Rohlf, 2007; Kaufmann et al., 2017; Kubinyi et al., 2009; Starling et al., 2013a). This suggests that the expression of behavior might also depend on the dog's reproductive status. The males in the current study were, with one exception, intact, and testing of neutered dogs may result in different findings.

4.7. Areas for future research

Social factors, environment, and experiences affect behavioral development (Foyer et al., 2013; Harvey et al., 2016a; Serpell & Duffy, 2016) Furthermore, the expression of personality traits may also depend on age, sex, and reproductive status (Bennett & Rohlf, 2007; Hart & Hart, 2017; Starling et al., 2013a), as well as owner characteristics (Kubinyi et al., 2009; Svartberg, 2002) and the specific context (Goold & Newberry, 2017a). Despite this, many studies on working dogs lack additional information about the dogs' behavior outside the test situation (e.g. Sinn et al., 2010; Wilsson & Sundgren, 1997). Furthermore, most police and military dog studies focus on predicting withdraw or completion of training (Foyer et al., 2014; Sinn et al., 2010; Slabbert & Odendaal, 1999; Wilsson & Sundgren, 1997; Wilsson & Sinn, 2012).

Longitudinal studies following dogs from puppies and into active service might provide useful information about personality traits, as well as environment factors, that contribute to suitable working dogs. However, a common problem with longitudinal studies is the loss of dogs between tests (Riemer et al., 2014; Svobodová et al., 2008), and an alternative could be to use questionnaires to collect data about early environment and behavior outside the test situation. Questionnaires are a useful tool when assessing personality traits of dogs (Wiener & Haskell, 2016), and previous studies have reported a good return rate (>70%) (Foyer et al., 2014; Svartberg, 2005), suggesting that the loss of subjects could be lower compared to repeated testing. Additionally, questionnaires have been shown to provide valuable information about which external factors can affect test performance (e.g. Fuchs et al., 2005). Fuchs and his team (2005) discovered that external factors affected German shepherds' test performance in a later test (average age 20 and 29 months, respectively). Dogs with regular contact with humans (adults and school age children) showed better defense drive, while dogs lacking young dog training exhibited less confidence and nerve stability (Fuchs et al., 2005). Furthermore, questionnaires have been used to predict future test outcome in service and working dogs (Duffy & Serpell, 2012; Foyer et al., 2014).

A future study combining questionnaire data with test data would, therefore, be useful for assessing how owner reports on behavior outside the test context could contribute to the assessment of dog suitability for police work. Additionally, given that the current findings are restricted to intact male German shepherds, it would be relevant to build on the current findings by evaluating earlier testing of females and additional breeds.

4.8. Conclusions

My results give evidence supporting earlier assessment of police patrol dogs than is currently practiced. I found that test outcome (pass vs fail) at 6 months had moderate predictive validity for the test outcome at 12 months of age. Furthermore, seven subtests had an average score that was consistent in the interval 6-12 months. These subtests describe behaviors that are considered desirable in both police and military dogs (Wilsson & Sinn, 2012), and the results of this study can be useful for further studies on the selection of working dogs.

In the current study, dogs that passed or failed the qualification test at 6 months with high or low behavior scores, respectively, received the same test outcome at 12 months, whereas dogs with ambiguous (intermediate) behavior scores at 6 months scored higher at 12 months of age (i.e. passed the qualification test at 12 months, after having failed at 6 months). Thus, I suggest a primarily test for police patrol dog suitability when the dogs are 6 months of age to exclude low scoring dogs and accept high scoring dogs for further training. Dogs with intermediate results at 6 months could be re-tested at 12 months of age to minimize the loss of suitable dogs. The implementation of testing police patrol dogs at 6 months, with the possibility of a second test at 12 months, would increase dog welfare by allowing unsuited dogs to be re-homed earlier. Furthermore, it would reduce rearing and training cost without the loss of potential police dogs.

6. References

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London: Chapman and Hall.

Bennett, P. C. & Rohlf, V. I. (2007). Owner-companion dog interactions: Relationships between demographic variables, potentially problematic behaviours, training engagement and shared activities. *Applied Animal Behaviour Science*, 102 (1): 65-84. doi: https://doi.org/10.1016/j.applanim.2006.03.009.

Bland, J. M. & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8 (2): 135-160. doi: https://doi.org/10.1177/096228029900800204.

Bradshaw, J. W. S., Pullen, A. J. & Rooney, N. J. (2015). Why do adult dogs 'play'? *Behavioural Processes*, 110: 82-87. doi: https://doi.org/10.1016/j.beproc.2014.09.023.

Careau, V., Réale, D., Humphries, Murray M. & Thomas, Donald W. (2010). The Pace of Life under Artificial Selection: Personality, Energy Expenditure, and Longevity Are Correlated in Domestic Dogs. *The American Naturalist*, 175 (6): 753-758. doi: https://doi.org/10.1086/652435.

Chaloupková, H., Svobodová, I., Vápeník, P. & Bartoš, L. (2018). Increased resistance to sudden noise by audio stimulation during early ontogeny in German shepherd puppies. *PLOS ONE*, 13 (5): e0196553. doi: https://doi.org/10.1371/journal.pone.0196553.

Cobb, M., Branson, N., McGreevy, P., Lill, A. & Bennett, P. (2015). The advent of canine performance science: Offering a sustainable future for working dogs. *Behavioural Processes*, 110: 96-104. doi: https://doi.org/10.1016/j.beproc.2014.10.012.

Diederich, C. & Giffroy, J.-M. (2006). Behavioural testing in dogs: A review of methodology in search for standardisation. *Applied Animal Behaviour Science*, 97 (1): 51-72. doi: https://doi.org/10.1016/j.applanim.2005.11.018.

Duffy, D. L. & Serpell, J. A. (2012). Predictive validity of a method for evaluating temperament in young guide and service dogs. *Applied Animal Behaviour Science*, 138 (1): 99-109. doi: https://doi.org/10.1016/j.applanim.2012.02.011.

Fleiss, J. L., Levin, B. & Paik, M. C. (2003). *Statistical methods for rates and proportions*. 3rd ed. Hoboken: John Wiley & Sons.

Foyer, P., Wilsson, E., Wright, D. & Jensen, P. (2013). Early experiences modulate stress coping in a population of German shepherd dogs. *Applied Animal Behaviour Science*, 146 (1): 79-87. doi: https://doi.org/10.1016/j.applanim.2013.03.013.

Foyer, P., Bjällerhag, N., Wilsson, E. & Jensen, P. (2014). Behaviour and experiences of dogs during the first year of life predict the outcome in a later temperament test. *Applied Animal Behaviour Science*, 155: 93-100. doi: https://doi.org/10.1016/j.applanim.2014.03.006.

Fratkin, J. L., Sinn, D. L., Patall, E. A. & Gosling, S. D. (2013). Personality Consistency in Dogs: A Meta-Analysis. *PLOS ONE*, 8 (1): e54907. doi: https://doi.org/10.1371/journal.pone.0054907.

Fratkin, J. L., Sinn, D. L., Thomas, S., Hilliard, S., Olson, Z. & Gosling, S. D. (2015). Do you see what I see? Can non-experts with minimal training reproduce expert ratings in behavioral assessments of working dogs? *Behavioural Processes*, 110: 105-116. doi: https://doi.org/10.1016/j.beproc.2014.09.028.

Fuchs, T., Gaillard, C., Gebhardt-Henrich, S., Ruefenacht, S. & Steiger, A. (2005). External factors and reproducibility of the behaviour test in German shepherd dogs in Switzerland. *Applied Animal Behaviour Science*, 94 (3): 287-301. doi: https://doi.org/10.1016/j.applanim.2005.02.016.

Gamer, M., Lemon, J. & Singh, I. F. P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (Version 0.84.1). Available at: https://CRAN.R-project.org/package=irr.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25 (2): 141-151. doi: https://doi.org/10.11613/BM.2015.015.

Goddard, M. E. & Beilharz, R. G. (1984). A factor analysis of fearfulness in potential guide dogs. *Applied Animal Behaviour Science*, 12 (3): 253-265. doi: https://doi.org/10.1016/0168-1591(84)90118-7.

Goddard, M. E. & Beilharz, R. G. (1986). Early prediction of adult behaviour in potential guide dogs. *Applied Animal Behaviour Science*, 15 (3): 247-260. doi: https://doi.org/10.1016/0168-1591(86)90095-X.

Goold, C., Vas, J., Olsen, C. & Newberry, R. C. (2016). Using network analysis to study behavioural phenotypes: an example using domestic dogs. *Royal Society Open Science*, 3 (10): 160268. doi: https://doi.org/10.1098/rsos.160268.

Goold, C. & Newberry, R. C. (2017a). Aggressiveness as a latent personality trait of domestic dogs: Testing local independence and measurement invariance. *PLOS ONE*, 12 (8): e0183595. doi: https://doi.org/10.1371/journal.pone.0183595.

Goold, C. & Newberry, R. C. (2017b). Modelling personality, plasticity and predictability in shelter dogs. *Royal Society Open Science*, 4 (9): 170618. doi: https://doi.org/10.1098/rsos.170618.

Gray, J. A. (1987). *The psychology of fear and stress*. 2nd ed. Cambridge: Cambridge University Press.

Han, J., Kamber, M. & Pei, J. (2012). *Data Mining: Concepts and Techniques*. 3rd ed. Boston: Morgan Kaufmann.

Hart, B. L. & Hart, L. A. (2017). Breed and gender differences in dog behavior. In Serpell, J. A. (ed.) *The domestic dog: its evolution, behavior and interactions with people*, pp. 118-132. Cambridge: Cambridge University Press.

Harvey, N. D., Craigon, P. J., Blythe, S. A., England, G. C. W. & Asher, L. (2016a). Social rearing environment influences dog behavioral development. *Journal of Veterinary Behavior*, 16: 13-21. doi: https://doi.org/10.1016/j.jveb.2016.03.004.

Harvey, N. D., Craigon, P. J., Sommerville, R., McMillan, C., Green, M., England, G. C. W. & Asher, L. (2016b). Test-retest reliability and predictive validity of a juvenile guide dog behavior test. *Journal of Veterinary Behavior*, 11: 65-76. doi: https://doi.org/10.1016/j.jveb.2015.09.005.

Hoummady, S., Péron, F., Grandjean, D., Cléro, D., Bernard, B., Titeux, E., Desquilbet, L. & Gilbert, C. (2016). Relationships between personality of human–dog dyads and performances in working tasks. *Applied Animal Behaviour Science*, 177: 42-51. doi: https://doi.org/10.1016/j.applanim.2016.01.015.

Jamieson, L. T. J., Baxter, G. S. & Murray, P. J. (2017). Identifying suitable detection dogs. *Applied Animal Behaviour Science*, 195: 1-7. doi: https://doi.org/10.1016/j.applanim.2017.06.010.

Kalra, A. (2017). Decoding the Bland–Altman plot: Basic review. *Journal of the Practice of Cardiovascular Sciences*, 3 (1): 36-38. doi: http://dx.doi.org/10.4103/jpcs.jpcs_11_17.

Kaufmann, C. A., Forndran, S., Stauber, C., Woerner, K. & Gansloßer, U. (2017). The social behaviour of neutered male dogs compared to intact dogs (Canis lupus familiaris): Video analyses, questionnaires and case studies. *Veterinary Medicine Open Journal*, 2 (1): 22-37. doi: http://dx.doi.org/10.17140/VMOJ-2-113.

Kubinyi, E., Turcsán, B. & Miklósi, Á. (2009). Dog and owner demographic characteristics and dog personality trait associations. *Behavioural Processes*, 81 (3): 392-401. doi: https://doi.org/10.1016/j.beproc.2009.04.004.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28 (5): 1-26. doi: http://dx.doi.org/10.18637/jss.v028.i05.

LaMorte, W. W. (2016). *Positive and Negative Predictive Value*. Available at: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Screening/EP713_Screening5.html (accessed: 15.04.2019).

Lehner, P. N. (1996). *Handbook of ethological methods*. 2nd ed. Cambridge: Cambridge University Press.

Lehnert, B. (2015). *BlandAltmanLeh: Plots (Slightly Extended) Bland-Altman Plots* (Version 0.3.1). Available at: https://CRAN.R-project.org/package=BlandAltmanLeh.

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.3.2). Available at: https://CRAN.R-project.org/package=emmeans.

Li, J., Cheng, K. & Liu, W. (2019). *StepReg: Stepwise Regression Analysis* (Version 1.0.1). Available at: https://cran.r-project.org/web/packages/StepReg/index.html.

Lin, W.-L. & Yao, G. (2014). Predictive Validity. In Michalos, A. C. (ed.) *Encyclopedia of Quality of Life and Well-Being Research*, pp. 5020-5021. Dordrecht: Springer Netherlands.

Lord, K., Coppinger, L. & Coppinger, R. (2014). Differences in the behavior of landraces and breeds of dogs. In Grandin, T. & Jessinger, M. J. (eds) *Genetics and the behavior of domesticated animals* pp. 195-235. San Diego: Academic Press, Elsevier Inc.

Lord, K., Schneider, R. A. & Coppinger, R. (2017). Evolution of working dogs. In Serpell, J. A. (ed.) *The domestic dog: its evolution, behavior and interactions with people*, pp. 42-66. Cambridge Cambridge University Press.

Martin, P. & Bateson, P. (2007). *Measuring Behaviour: An Introductory Guide*. 3rd ed. Cambridge: Cambridge University Press.

McFarland, D. (2006). *Dictionary of Animal Behaviour*. Oxford: Oxford University Oress.

Miklósi, Á. (2015). *Dog behaviour, evolution, and cognition*. 2nd ed. Oxford: Oxford University Press.

Overall, K. L. (2013). *Manual of Clinical Behavioral Medicine for Dogs and Cats*. St Louis: Mosby Elsevier Inc.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56 (1): 45-50. doi: http://dx.doi.org/10.4103/0301-4738.37595.

Patronek, G. J., Bradley, J. & Arps, E. (2019). What is the evidence for reliability and validity of behavior evaluations for shelter dogs? A prequel to “No better than flipping a coin”.

Journal of Veterinary Behavior, 31: 43-58. doi: https://doi.org/10.1016/j.jveb.2019.03.001.

R Core Team. (2018). *R: Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for Statistical Computing. Available at: https://www.r-project.org/.

Riemer, S., Müller, C., Virányi, Z., Huber, L. & Range, F. (2014). The Predictive Value of Early Behavioural Assessments in Pet Dogs – A Longitudinal Study from Neonates to Adults. *PLOS ONE*, 9 (7): e101237. doi: https://doi.org/10.1371/journal.pone.0101237v.

Rooney, N. J., Clark, C. C. A. & Casey, R. A. (2016). Minimizing fear and anxiety in working dogs: A review. *Journal of Veterinary Behavior*, 16: 53-64. doi: https://doi.org/10.1016/j.jveb.2016.11.001.

Serpell, J. A. & Hsu, Y. (2001). Development and validation of a novel method for evaluating behavior and temperament in guide dogs. *Applied Animal Behaviour Science*, 72 (4): 347-364. doi: https://doi.org/10.1016/S0168-1591(00)00210-0.

Serpell, J. A. & Hsu, Y. A. (2005). Effects of breed, sex, and neuter status on trainability in dogs. *Anthrozoös*, 18 (3): 196-207. doi: https://doi.org/10.2752/089279305785594135.

Serpell, J. A. & Duffy, D. L. (2016). Aspects of Juvenile and Adolescent Environment Predict Aggression and Fear in 12-Month-Old Guide Dogs. *Frontiers in Veterinary Science*, 3 (49). doi: https://doi.org/10.3389/fvets.2016.00049.

Serpell, J. A., Duffy, D. L. & Jagoe, J. A. (2017). Becoming a dog: Early experiences and the development of behavior. In Serpell, J. A. (ed.) *The domestic dog: its evolution, behavior and interactions with people*, pp. 93-117. Cambridge: Cambridge University Press.

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): 51-65. doi: https://doi.org/10.1016/j.applanim.2010.08.007.

Slabbert, J. M. & Odendaal, J. S. J. (1999). Early prediction of adult police dog efficiency—a longitudinal study. *Applied Animal Behaviour Science*, 64 (4): 269-288. doi: https://doi.org/10.1016/S0168-1591(99)00038-6.

Stamps, J. & Groothuis, T. G. G. (2010). The development of animal personality: relevance, concepts and perspectives. *Biological Reviews*, 85 (2): 301-325. doi: https://doi.org/10.1111/j.1469-185X.2009.00103.x.

Starling, M. J., Branson, N., Thomson, P. C. & McGreevy, P. D. (2013a). Age, sex and reproductive status affect boldness in dogs. *The Veterinary Journal*, 197 (3): 868-872. doi: https://doi.org/10.1016/j.tvjl.2013.05.019.

Starling, M. J., Branson, N., Thomson, P. C. & McGreevy, P. D. (2013b). "Boldness" in the domestic dog differs among breeds and breed groups. *Behavioural Processes*, 97: 53-62. doi: https://doi.org/10.1016/j.beproc.2013.04.008.

Svartberg, K. (2002). Shyness–boldness predicts performance in working dogs. *Applied Animal Behaviour Science*, 79 (2): 157-174. doi: https://doi.org/10.1016/S0168-1591(02)00120-X.

Svartberg, K. & Forkman, B. (2002). Personality traits in the domestic dog (Canis familiaris). *Applied Animal Behaviour Science*, 79 (2): 133-155. doi: https://doi.org/10.1016/S0168-1591(02)00121-1.

Svartberg, K. (2005). A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. *Applied Animal Behaviour Science*, 91 (1): 103-128. doi: https://doi.org/10.1016/j.applanim.2004.08.030.

Svartberg, K., Tapper, I., Temrin, H., Radesäter, T. & Thorman, S. (2005). Consistency of personality traits in dogs. *Animal Behaviour*, 69 (2): 283-291. doi: https://doi.org/10.1016/j.anbehav.2004.04.011.

Svobodová, I., Vápeník, P., Pinc, L. & Bartoš, L. (2008). Testing German shepherd puppies to assess their chances of certification. *Applied Animal Behaviour Science*, 113 (1): 139-149. doi: https://doi.org/10.1016/j.applanim.2007.09.010.

Taylor, K. D. & Mills, D. S. (2006). The development and assessment of temperament tests for adult companion dogs. *Journal of Veterinary Behavior*, 1 (3): 94-108. doi: https://doi.org/10.1016/j.jveb.2006.09.002.

van den Berg, L. (2017). Genetics of dog behavior. In Serpell, J. A. (ed.) *The domestic dog: its evolution, behavior and interactions with people*, pp. 69-92. Cambridge: Cambridge University Press.

Vas, J., Topál, J., Péch, É. & Miklósi, Á. (2007). Measuring attention deficit and activity in dogs: A new application and validation of a human ADHD questionnaire. *Applied Animal Behaviour Science*, 103 (1): 105-117. doi: https://doi.org/10.1016/j.applanim.2006.03.017.

Whitlock, M. C. & Schluter, D. (2015). *The analysis of biological data*. 2nd ed. Greenwood Village: Roberts and Company Publishers.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. New York: Springer-Verlag.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'* (Version 1.2.1). Available at: https://CRAN.R-project.org/package=tidyverse.

Wiener, P. & Haskell, M. J. (2016). Use of questionnaire-based data to assess dog personality. *Journal of Veterinary Behavior*, 16: 81-85. doi: https://doi.org/10.1016/j.jveb.2016.10.007.

Wilsson, E. & Sundgren, P.-E. (1997). The use of a 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 dogs, sex and breed differences. *Applied Animal Behaviour Science*, 53 (4): 279-295. doi: https://doi.org/10.1016/S0168-1591(96)01174-4.

Wilsson, E. & Sundgren, P.-E. (1998). Behaviour test for eight-week old puppies—heritabilities of tested behaviour traits and its correspondence to later behaviour. *Applied Animal Behaviour Science*, 58 (1): 151-162. doi: https://doi.org/10.1016/S0168-1591(97)00093-2.

Wilsson, E. & Sinn, D. L. (2012). Are there differences between behavioral measurement methods? A comparison of the predictive validity of two ratings methods in a working dog program. *Applied Animal Behaviour Science*, 141 (3): 158-172. doi: https://doi.org/10.1016/j.applanim.2012.08.012.

Appendices

Appendix 1 – Test redundancy

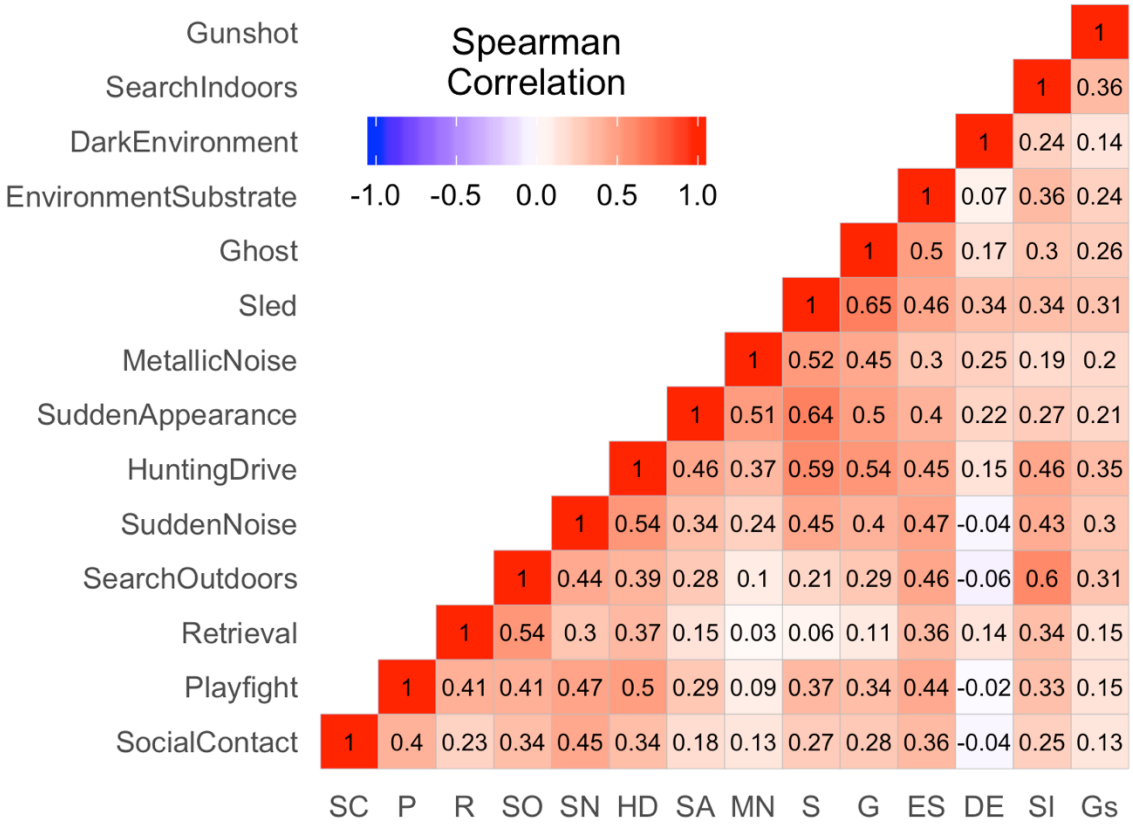

Figure A1. Heat map illustrating the correlation (i.e. redundancy) between the average scores for the 14 subtests: Social contact (SC), Playfight (P), Retrieval (R), Search outdoors (SO), Sudden noise (SN), Hunting Drive (HD), Sudden appearance (SA), Metallic noise (MN), Sled (S), Ghost (G), Environment substrate (ES), Dark environment (DE), Search indoors (SI), and Gunshot (Gs). The heat map is based on a Spearman correlation matrix (N = 62), and the numbers represent the Spearman correlation coefficients.

Appendix 2 – Average subtest scores representing behavioral variables

Table A2. Difference in scoring between the dogs (N = 62) at 6 months and 12 months of age in the 14 subtests. Means, standard deviations and p-values are given for the subtest score and the corresponding variables. Wilcoxon matched paired-test requires an equal number of individuals in both groups, and in cases of missing values in one age group, the corresponding dog was removed from the other age group. This is indicated by the number behind the variable name. V represents the test statistic for Wilcoxon matched paired-test, and *p*-values < 0.05 are in bold.

Behavioral variables	6 months	12 months	Statistics	
	Mean ± SD	Mean ± SD	V	p-value
Social contact	4.15 ± 0.64	4.34 ± 0.59	490.5	**0.047**
Contact with strangers	4.50 ± 0.95	3.65 ± 0.68	50.0	0.208
Social confidence	4.16 ± 1.03	4.40 ± 0.86	58.0	**0.040**
Contact with tester	4.42 ± 0.92	4.53 ± 0.84	86.5	0.479
Following	3.95 ± 0.89	3.94 ± 0.92	254.5	0.904
Handling	4.26 ± 0.87	4.39 ± 0.96	224.5	0.452
Confidence (N = 55)	3.58 ± 0.88	4.24 ± 0.94	81.0	**<0.001**
Playfight	3.81 ± 0.50	3.95 ± 0.60	433.5	**0.032**
Intensity	3.35 ± 0.79	3.42 ± 0.90	164.5	0.528
Grip strength	3.87 ± 0.97	3.81 ± 1.05	332.0	0.535
Drive	3.74 ± 0.87	3.85 ± 0.96	79.0	0.323
Resilience	3.87 ± 0.53	3.90 ± 0.76	126.5	0.708
Aggression	4.37 ± 0.94	4.35 ± 0.83	183.0	0.853
Confidence	3.63 ± 0.89	4.34 ± 0.87	86.5	**<0.001**
Retrieval	3.94 ± 0.52	4.26 ± 0.52	230.0	**0.001**
Cooperation	3.52 ± 1.18	3.84 ± 1.10	118.0	**0.016**
Aggression	4.58 ± 0.69	4.42 ± 0.80	98.0	0.114
Confidence (N = 61)	3.72 ± 0.93	4.48 ± 0.79	52.0	**<0.001**
Search outdoors[1]	3.79 ± 1.10	4.19 ± 0.82	279.0	**0.003**
Focus	3.81 ± 1.34	4.23 ± 0.93	190.5	**0.012**
Tracking ability	4.05 ± 1.32	4.50 ± 1.02	66.0	**0.005**
Cooperation	3.52 ± 1.18	3.84 ± 1.10	160.5	**0.048**

Table A2 continued

Behavioral variables	6 months	12 months	Statistics	
	Mean ± SD	Mean ± SD	V	p-value
Sudden noise	3.94 ± 0.79	4.28 ± 0.72	272.0	**0.001**
Reaction to noise	4.15 ± 1.01	4.11 ± 1.01	265.5	0.985
Curiosity	3.32 ± 1.64	3.87 ± 1.60	119.0	**0.006**
Aggression	4.73 ± 0.58	4.69 ± 0.53	70.0	0.559
Confidence (N = 61)	3.54 ± 0.96	4.46 ± 0.92	11.0	**<0.001**
Hunting drive	3.35 ± 0.92	4.00 ± 0.95	270.0	**<0.001**
Intensity	3.00 ± 1.10	3.82 ± 1.18	153.5	**<0.001**
Interest	3.32 ± 1.35	4.00 ± 1.23	107.0	**<0.001**
Aggression	3.73 ± 1.07	4.11 ± 0.99	282.5	**0.031**
Confidence	3.35 ± 0.98	4.06 ± 1.13	90.0	**<0.001**
Sudden appearance	3.45 ± 0.67	3.47 ± 0.71	663.5	0.495
Startle response	3.55 ± 0.94	3.90 ± 0.84	156.0	**0.003**
Defense	3.79 ± 1.06	3.10 ± 1.14	735.5	**<0.001**
Exploration (N = 60)	2.72 ± 1.18	3.35 ± 1.29	151.0	**0.001**
Avoidance	3.74 ± 0.99	4.53 ± 1.20	440.5	0.168
Metallic noise	3.97 ± 0.70	3.99 ± 0.78	710.0	0.851
Startle response	3.58 ± 1.17	3.68 ± 1.00	307.0	0.480
Exploration	3.52 ± 1.05	3.79 ± 0.98	358.5	0.061
Aggression (N = 60)	4.42 ± 0.83	4.40 ± 0.98	172.5	0.948
Avoidance (N = 61)	4.43 ± 0.83	4.13 ± 1.06	437.0	**0.037**
Sled	3.31 ± 0.64	3.35 ± 0.77	819.0	0.620
Defense	3.66 ± 1.05	3.56 ± 1.28	328.0	0.595
Startle response	3.37 ± 1.07	3.66 ± 1.09	239.5	0.083
Threat response	3.81 ± 1.04	3.61 ± 1.21	406.5	0.240
Exploration	2.60 ± 1.06	3.02 ± 1.32	312.0	**0.029**
Avoidance	3.39 ± 1.30	3.29 ± 1.43	590.0	0.781
Aggression (N=60)	3.07 ± 0.99	3.03 ± 1.18	390.0	0.774

Table A2 continued

Behavioral variables	6 months	12 months	Statistics	
	Mean ± SD	Mean ± SD	V	p-value
Ghost	3.42 ± 0.71	3.54 ± 1.02	595.5	0.145
Threats	3.00 ± 1.48	3.21 ± 1.45	334.0	0.304
Startle response	3.95 ± 1.22	4.02 ± 1.30	390.0	0.592
Exploration	3.55 ± 1.28	3.89 ± 1.46	277.0	0.110
Aggression (N = 60)	3.23 ± 1.17	2.98 ± 1.17	489.5	0.072
Confidence (N=59)	3.46 ± 0.86	3.81 ± 1.20	165.5	0.054
Environment substrate	4.18 ± 0.50	4.19 ± 1.96	441.5	0.059
Aggression 1 (N = 55)	4.82 ± 0.43	4.65 ± 0.97	49.5	0.422
Confidence 1 (N = 61)	3.07 ± 1.63	3.69 ± 1.71	84.0	**0.001**
Aggression 2 (N = 60)	4.98 ± 0.13	4.75 ± 0.91	19.5	0.071
Confidence 2 (N = 60)	4.83 ± 0.67	4.63 ± 1.07	34.5	0.158
Play	3.37 ± 1.01	3.42 ± 1.02	216.0	0.733
Dark environment	4.39 ± 0.83	4.44 ± 1.03	301.5	0.319
Aggression (N = 61)	4.66 ± 0.66	4.48 ± 1.01	139.5	0.192
Confidence (N = 61)	4.38 ± 1.19	4.64 ± 1.07	64.0	0.198
Curiosity	4.13 ± 1.25	4.32 ± 1.11	231.5	0.250
Search indoors[1]	3.64 ± 1.08	4.08 ± 1.16	355.0	**0.001**
Focus	3.44 ± 1.46	4.18 ± 1.32	194.0	**0.001**
Tracking ability	3.73 ± 1.33	4.16 ± 1.28	294.0	**0.017**
Cooperation	3.76 ± 1.21	3.89 ± 1.28	210.5	0.455
Gunshot	4.40 ± 0.48	4.30 ± 0.94	711.5	0.789
Reaction	4.29 ± 0.52	3.92 ± 0.84	331.0	**0.002**
Aggression (N = 60)	4.80 ± 0.44	4.45 ± 0.95	203.0	**0.007**
Confidence (N = 61)	4.13 ± 0.85	4.33 ± 0.98	243.0	0.149
Reaction during play	4.79 ± 0.52	4.48 ± 1.17	106.0	**0.048**
Confidence (N = 60)	4.08 ± 0.94	4.50 ± 1.00	207.5	**0.009**

[1] Time in search was measured on a different scale (i.e. in seconds), and was not included in the calculation of the subtest scores of Search outdoors and Search indoors.

Appendix 3 – Factor analysis

Appendix 3.1. Scree test

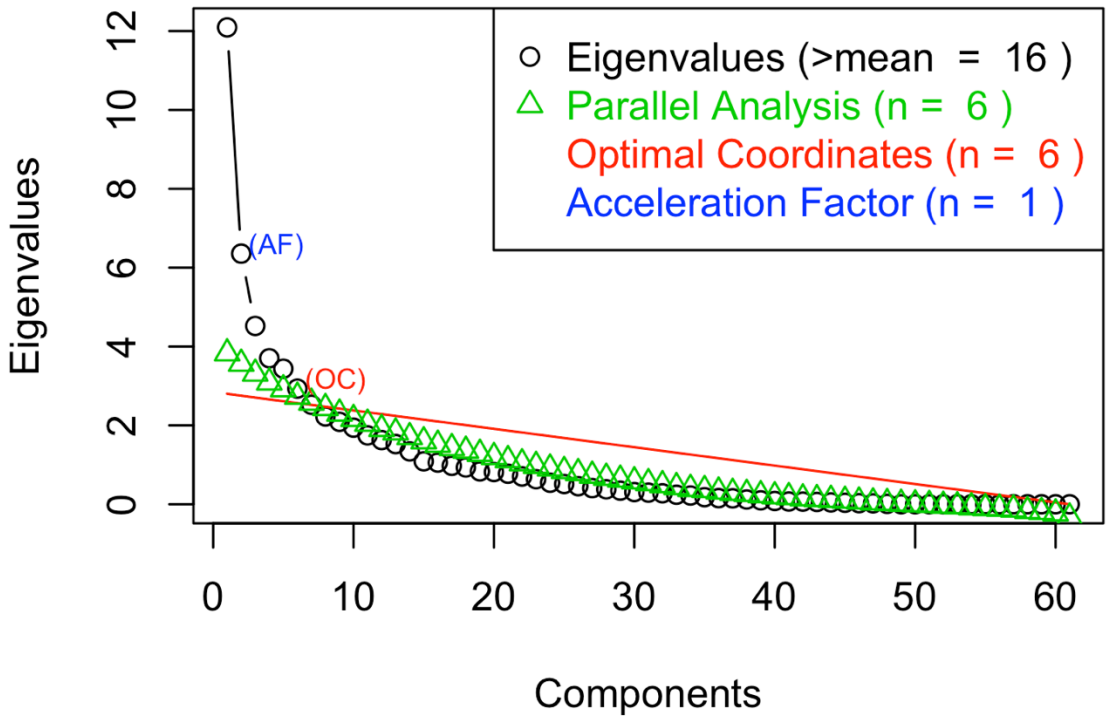

Figure A3. A graphical solution to the Scree test. The y-axis shows the eigenvalues and the number of components (i.e. factors) is on the x-axis. This Scree test identifying 6 factors as the optimal number of factors to be retained in the factor analysis based on the number of observations from a Spearman correlation matrix (N = 49).

Appendix 3.2. Factor analysis

Table A3. Three of the six extracted factors were similar to Svartberg and Forkman (2002), and are listed as 'Identified factor'. The remaining three factors are listed under 'Other factors'. Loadings ≥ 40 are in bold. Variables used to calculate the factor scores are in bold, and their loading on the respective factor scores are underlined.

Subtest	Variable	Identified factors			Other factors		
		Playfulness	Sociability	Curiosity/Fearlessness	Factor 1	Factor 2	Factor 3
Social contact	**Contact with strangers**	–0.04	**0.53**	–0.12	–0.03	–0.05	0.01
	Social confidence	0.01	**0.69**	–0.01	0.02	0.02	0.03
	Contact with tester	0.06	**0.46**	–0.17	–0.18	0.20	–0.02
	Following	0.10	**0.71**	0.15	0.12	0.02	–0.09
	Handling	0.04	**0.58**	–0.06	0.22	0.06	–0.01
	Confidence	0.14	**0.48**	–0.07	–0.38	–0.19	0.19
Playfight	**Intensity**	**0.73**	0.06	–0.01	0.02	0.03	–0.02
	Grip strength	**0.65**	–0.01	0.20	–0.02	–0.15	0.05
	Drive	**0.73**	0.07	–0.02	–0.06	0.03	0.01
	Resilience	**0.64**	0.09	–0.04	0.04	–0.01	0.08
	Aggression	–0.18	0.18	0.00	**0.84**	–0.12	0.05
	Confidence	0.17	0.26	–0.12	**–0.40**	–0.30	0.12
Retrieval	Cooperation	–0.15	–0.03	–0.01	0.06	–0.01	**0.90**
	Aggression	–0.10	0.04	–0.06	**0.89**	–0.05	0.05
	Confidence	0.29	0.29	–0.08	–0.34	–0.15	0.25
Search outdoors	Focus	**0.51**	0.06	0.01	–0.07	–0.24	**0.48**
	Tracking ability	**0.52**	–0.28	0.08	–0.17	0.00	**0.40**
	Cooperation	0.04	–0.03	–0.08	–0.05	0.14	**0.93**

Table 3A continued

Subtest	Variable	Identified factors			Other factors		
		Playfulness	Sociability	Curiosity/Fearlessness	Factor 1	Factor 2	Factor 3
Sudden noise	Reaction to noise	0.17	0.38	0.30	0.01	–0.16	0.21
	Curiosity	0.36	0.30	0.22	0.19	–0.16	0.11
	Aggression	0.24	0.10	0.07	**0.70**	0.02	0.01
	Confidence	**0.62**	0.23	0.08	–0.11	–0.19	0.07
Hunting drive	Intensity	0.35	0.12	**0.47**	0.00	–0.12	–0.09
	Interest	**0.44**	0.15	**0.52**	0.02	–0.13	–0.13
	Aggression	–0.06	0.02	**0.65**	0.18	–0.24	–0.03
	Confidence	0.16	0.26	0.33	–0.20	–0.27	0.13
Sudden appearance	**Startle response**	0.05	0.13	**0.50**	–0.28	–0.21	–0.13
	Defense	0.02	0.21	**0.40**	0.07	–0.01	0.22
	Exploration	0.18	0.00	0.37	0.06	0.04	–0.05
	Avoidance	0.05	**0.58**	0.18	0.24	0.13	0.14
Metallic noise	**Startle response**	–0.01	–0.06	**0.57**	–0.30	–0.02	–0.10
	Exploration	0.29	–0.27	**0.55**	0.08	0.15	0.00
	Aggression	–0.15	0.06	0.17	0.36	0.14	0.23
	Avoidance	0.09	0.28	**0.50**	0.14	0.22	0.03
Sled	Defense	0.05	–0.03	**0.41**	0.36	0.33	0.06
	Startle response	0.18	0.30	0.31	0.12	0.29	–0.09
	Threat response	–0.03	–0.30	0.34	–0.07	**0.45**	0.12
	Exploration	0.32	0.26	0.29	–0.11	0.21	–0.02

Table 3A continued

Subtest	Variable	Identified factors			Other factors		
		Playfulness	Sociability	Curiosity/Fearlessness	Factor 1	Factor 2	Factor 3
	Avoidance	0.29	0.38	0.33	–0.08	0.04	–0.08
	Aggression	0.07	0.31	0.12	–0.25	0.07	–0.32
Ghost	Threats	0.27	–0.03	–0.02	0.37	**0.47**	–0.21
	Startle response	–0.24	0.20	**0.41**	–0.29	–0.10	0.13
	Exploration	–0.01	0.11	**0.57**	–0.18	0.10	0.09
	Aggression	–0.34	0.25	0.28	–0.38	–0.34	0.13
	Confidence	0.02	0.34	0.24	**–0.57**	0.14	0.29
Environment substrate	Aggression	0.05	–0.01	0.17	0.38	0.09	0.02
	Confidence	0.02	0.12	0.19	**–0.41**	–0.35	0.34
	Aggression	–0.23	–0.07	**0.45**	–0.11	0.33	–0.06
	Confidence	0.09	0.25	–0.23	–0.28	0.26	–0.18
	Play	**0.56**	0.18	–0.02	–0.21	**0.41**	–0.12
Dark environment	Aggression	–0.31	0.00	**0.41**	0.35	0.25	–0.03
	Confidence	–0.04	0.14	0.02	–0.02	**0.81**	0.07
	Curiosity	–0.02	–0.01	–0.06	–0.02	0.79	0.17
Search indoors	Focus	**0.53**	0.10	0.07	–0.07	0.20	0.11
	Tracking ability	**0.53**	–0.24	0.00	0.04	0.13	0.23
	Cooperation	0.21	0.09	0.02	0.23	0.09	**0.72**

Subtest	Variable	Identified factors			Other factors		
		Playfulness	Sociability	Curiosity/Fearlessness	Factor 1	Factor 2	Factor 3
Gunshot	Reaction	0.28	–0.39	0.24	0.00	0.16	0.05
	Aggression	0.20	–0.07	0.06	**0.59**	–0.07	–0.02
	Confidence	0.15	–0.20	0.15	–0.31	–0.33	0.06
	Reaction during play	0.02	0.16	0.16	0.05	–0.23	0.26
	Confidence	0.18	0.29	–0.13	**–0.47**	–0.32	0.10

The initial factor analysis was primarily conducted to investigate the possibility of calculating a Boldness score, and the results were not used in any analyses in this paper. The factor analysis was based on a Spearman correlation matrix (N = 49).

Appendix 4 – Temporal consistency: Selected variables score and Boldness score

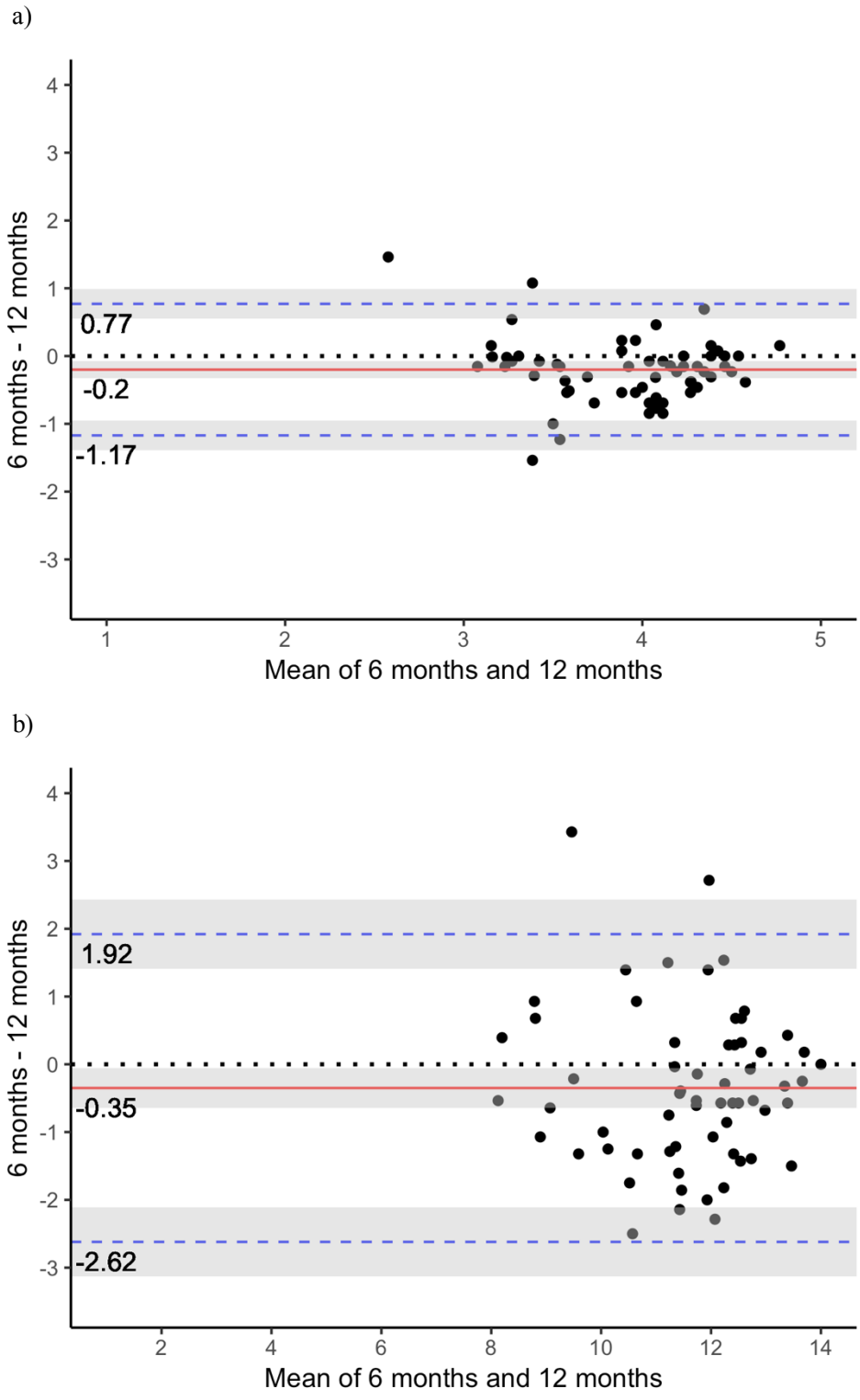

Figure A4. Bland-Altman plot of the a) Selected variables score and the b) Boldness score. The y-axis shows the score difference between the two test ages. The lines represent the mean difference (red solid line), the limits of agreement (blue dashed lines), and the point of zero difference (dotted line). Light grey areas present the 95% confidence interval for the mean difference and agreement limits.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway