

Validation of a method for assessing the ability of trainee specialist search dogs

Nicola Jane Rooney^{a,*}, Samantha Anne Gaines^{a,b},
John William Stephen Bradshaw^a, Stephen Penman^c

^a *Anthrozoology Institute, University of Bristol, Department of Clinical Veterinary Science, Langford, Bristol BS40 5DU, UK*

^b *Defence Science and Technology Laboratory, Fort Halstead, Sevenoaks, Kent TN14 7BP, UK*

^c *Defence Animal Centre, Melton Mowbray, Leicestershire LE13 0GX, UK*

Accepted 28 March 2006

Available online 11 May 2006

Abstract

To date, there are few validated tests for quantifying the ability of working dogs, and none documented for use on specialist search dogs. Such tests are essential to the empirical examination of ways to improve the efficiency of search dogs, a process critical to meet the increased demand for search dogs in a climate of global terrorist threat. This paper describes the development of a standard search task, which provides a systematic method for assessing the effectiveness of arms and explosives search dogs following training. This is the first documented use of ethological techniques to validate traditional ratings, which are based on handlers' opinions.

The subjects were 26 male Labrador retrievers, which had completed 10 weeks of standardised search training to enable them to indicate the presence of a range of target scents (explosives). At the end of the training period each dog's ability to complete a standardised search task was assessed. The efficiency with which the dog searched and located a range of target scents, along with their concurrent behaviour, was recorded on videotape and assessed via both subjective and objective measures. Subjective ratings made by scientists were very similar to those made by experienced trainers, and produced two uncorrelated factors; general search ability, and ability to work without false indications. Objective ethological assessment of the dogs' behaviour produced four measures; free search thoroughness, location ability, systematic search behaviour, and mean number of false indications.

The majority of these measures correlated significantly to the trainers' ratings of the dogs as taken throughout the training. The strongest correlation was with the subjective rating for general search ability. Of the four objective measures, free search thoroughness and location ability, combined with a smaller contribution from systematic search behaviour, also correlated strongly with the trainers' ratings of the same dogs. False indications were, however, apparently not reflected in the ratings of general search ability. This

* Corresponding author. Tel.: +44 117 928 9469; fax: +44 117 928 9582.

E-mail address: Nicola.Rooney@bristol.ac.uk (N.J. Rooney).

suggests that the standard search task provides similar, but more detailed information compared to the more traditional subjective method of assessing ability. We conclude that the standard search task method is a useful tool for future search dog research and ability assessments.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Specialist search dog; Assessment; Working dog; Detection dog; Validation; Dog behaviour; Internal validity; External validity; Reliability

1. Introduction

Specialist search dogs (also known as detection dogs) are trained by many law enforcement agencies throughout the world to locate explosives, weapons and drugs, as well as a range of other substances including human bodies, termites, accelerants, melanomas and screw worms (Gazit and Terkel, 2003). In the current climate of terrorist threat, the use of search dogs, and in particular those trained to detect arms and explosives, is especially important, and the demand for trained dogs continues to increase. As a result, there is a growing body of research examining ways of obtaining additional potential dogs for training (Rooney and Bradshaw, 2004; Rooney et al., 2004) and factors, which affect the ability of trained search dogs (e.g. Shivik, 2002; Williams and Johnston, 2002; Gazit and Terkel, 2003).

In order for such research to progress and alternative ways of producing high quality dogs to be compared, it is essential to be able to assess the ultimate ability of trained search dogs empirically. To date there is no fully validated method for assessing proficiency, and although individual agencies have their own procedures, there is currently no means to compare them. During standard training procedures, judgements as to the relative ability of dogs, and whether they should continue training or be rejected, are usually made by experienced trainers. This method has the advantage that the dogs' behaviour throughout training can be taken into consideration; however, due to their subjective nature, such assessments may be open to individual bias. An alternative method, which may prove useful during scientific trials, is to assess the dogs' ability during a single testing session. This provides a point-sample, which may be more objective. However, in order for a one-off test to be useful, it is essential that its outcome measures are representative of the dog's usual performance while searching. We would predict that it should therefore bear some resemblance to the opinions of experienced operatives and training staff with knowledge of the individual dog. The aim of this study is to devise such a test.

When a standardised test procedure is employed, there are two options as to how the dog's behaviour is measured and quantified. Firstly, there is the option to assess behaviour in terms of subjective scales, rated by people observing the actual test or a video recording. This has been the preferred option for proficiency, and suitability tests employed by several working dog organisations (e.g. Wilsson and Sundgren, 1997a,b; Goddard and Beilharz, 1982), and has the advantage that the observers can get an overall opinion of the dogs' ability. However several studies have indicated low inter-observer reliability for some traits (Goddard and Beilharz, 1982; Murphy, 1995). A second method of assessing behaviour during a test is to objectively measure specific aspects of the dogs' behaviour, as has been employed in cross matching scent tasks (e.g. Schoon and De Bruin, 1994). This may be advantageous when examining specific abilities, to enable individual dimensions of performance to be considered in isolation. Svartberg et al. (2005) have developed scales to describe dogs' reactions to stimuli during a "personality test". These are rated by assessors but are based on clear behavioural descriptors and as such aim to

minimise raters' subjectivity. However, specific measures to assess search-related behaviour have never been documented. In this study, we develop a set of objective measures, which together describe all aspects of search performance, and we compare them to scales derived from subjective assessments. The reliability of subjective assessments relies upon agreement between experts who may vary in their experience of the attributes they are being asked to assess. In this study we make a comparison between ratings made by dog trainers and those made by scientists familiar with search work.

For a relatively straight-forward task, such as a dog's ability to recognise and cross-match a scent (e.g. Schoon, 1996, 1998; Schoon and De Bruin, 1994), ability can be easily assessed using a single pass/fail outcome measure. However, the role of a specialist search dog requires a range of abilities and skills (Rooney et al., 2004). Any assessment of the ability of search dogs must take all aspects of the role into consideration, and fully reflect the task which the dog will be required to perform. A search dog must locate target substances using its sense of smell. It is often required to search potentially large areas and continue searching for long periods with no appreciable decline in work rate. Those dogs trained to search for explosives fulfil a potentially dangerous role, and are thus required to search thoroughly under the distant control of their handler.

In UK, the majority of search work is carried out off lead. When the dog-handler team enter each area, which they are required to search, the dog enters first and searches independently; this is termed the "free search". A proficient dog will cover the majority of the area during this stage. Next, the dog is brought under closer control by the handler, who encourages it to search the entire area during a "systematic search". Ideally the dog will work ahead of its handler, who attracts the dog's attention to any features which it may have omitted to search. When the dog locates the target substance, it must alert the handler and indicate the position of the find. This can be done in several ways, but, increasingly in the UK, and exclusively within the subjects of this study, dogs are trained to show a "passive response"; they change posture, sitting or lying down and staring at the target substance. The dog remains in this position, close to and orientated towards, but not in contact with, the target. The assessment procedure described in this paper aims to assess all these aspects of the search task. Search dogs are required to search outside areas, disused buildings and roads, as well as indoor areas. In order to standardise the environment, it was decided that our test procedure should take place exclusively indoors.

We devised several assessment methods and compared their respective outcome measures. The ultimate ability of each dog was assessed via two different means; trainers' assessments, and a standard video-recorded assessment after 10 weeks of training. From the latter, two sets of measures were derived: subjective measures of ability, made by several dog trainers and scientists, and objective measures of behaviour, made by one ethologist. Thus, we were able to compare the ability of scientists to that of experienced dog trainers, when assessing trained search dogs. We were also able to examine the value of an independent objective assessment method and compare it to more conventional subjective techniques.

2. Methods

2.1. Subjects

The subjects were 26 entire male Labrador retrievers, which had all been purchased from breeders at the age of 8 weeks, and raised by volunteer puppy-walkers, under close supervision for the subsequent 9 months.

Subjects were removed from their puppy walking home at the age of 11–12 months, and entered a military dog training school. Each dog underwent a quarantine period of approximately 2 weeks (mean = 17 ± 5 days) before commencing training. Since the dogs' births spanned a 6-month period, they were trained in four separate groups; groups A–D contained six, nine, five and six dogs, respectively. The dogs within each group were a maximum of 32 days different in age, so they ranged from 352 to 390 days old when training commenced.

2.2. Training

All training was carried out at the Defence Animal Centre (DAC), which procures and trains dogs, including search dogs, for the Ministry of Defence. The dogs were trained by four military trainers working in pairs, balanced for training experience, one male and one female in each. The dogs were allocated to four groups (ABCD); group A and group C were trained by trainer pair 1 (one of which, SP, also supervised all training), whilst groups B and D were trained by trainer pair 2. Each individual dog was handled by one of the trainers, who had principal responsibility for that dog. Throughout training the trainers remained blind to the experimental design and knew nothing about the history of the dogs.

The trainers aimed to train all dogs, for the entire training period, no matter how poor the dog's aptitude or ability appeared to be. At the beginning of training there were 29 dogs in the sample, however due to welfare considerations, and/or exceptionally poor progress which hindered the training of the rest of the group, three of the initial dogs were eliminated before training was completed. These dogs are not included in any of the analysis or sample details presented.

Training of all the dogs followed a standard protocol, which included objectives for each of the 10 weeks of training. The protocol stipulated how many target scents the dogs were trained to find, and at what stage each target was introduced. The number of training sessions per day and the average time searching were also standardised, as were the types of areas in which the dogs were trained each week. The dogs were all trained for 3, 4 or 5 days each week. If, during a given week, the pair of trainers were unable to complete three training days, this was not classified as a training week and the week's training objectives were carried over to the next week. Thus, some of the dogs took more than 10 weeks to complete their training and their final assessment took place at a time appropriate to their individual training stage.

2.3. Trainers' assessments

2.3.1. Methods

For each training group, both trainers rated each of the dogs (both those they handled themselves and those their training partner handled), at the end of weeks 1, 4, 7 and 10 of training. In the case of the week 10 ratings, the written assessment was done before the dog performed the video-recorded assessment (see Section 2.4). This ensured that the trainers were rating their general impression of the dog's ability and not being affected by their performance during the assessment.

On each occasion, the trainers rated each dog for 12 different attributes. These were characteristics which previous research had found to be important in potential search dogs (Rooney and Bradshaw, 2004; Rooney et al., 2004, Table 1). For each characteristic the dog was rated as either: 1, extremely low; 2, low; 3, intermediate; 4, high; or 5, extremely high.

The trainers were also asked to rate the dog's overall ability on a scale of 1–5: 1, one of the worst dogs I have ever trained; 2, below average; 3, about average; 4, above average; 5, one of the best dogs I have ever trained.

During the final assessment during the 10th week of training, additional components were included. Trainers were asked whether, at the end of the course, they thought the dog: 4, would definitely be ready to progress on to a specialist training course; 3, would need some extra training and then be ready for specialist training; 2, might possibly be ready with some additional training; 1, should be rejected.

This formed their future fate (FF) rating.

Table 1
Characteristics rated by trainers following 1, 4, 7 and 10 weeks of training

| No. | Characteristic |
|-----|--|
| 1 | Obedience to human command |
| 2 | Boldness |
| 3 | Playfulness |
| 4 | Tendency to hunt by smell alone |
| 5 | Level of aggression towards humans |
| 6 | Stamina |
| 7 | Ability to learn from being rewarded |
| 8 | Interest in toys or objects |
| 9 | Acuity of sense of smell |
| 10 | Motivation to retain possession of an object |
| 11 | Health |
| 12 | Ease of adaptation to kennel environment |
| 13 | Overall ability |

Finally, in the 10th week of their first group of dogs, each trainer was asked to rate the level of each of the 12 attributes (Table 1) which they thought was ideally suited for an arms and explosives search dog. These ratings were used to calculate discrepancies from ideal for each dog, a technique developed for use on companion dogs (Serpell, 1996), and previously employed on search dogs (Rooney et al., 2004).

2.3.2. Deriving trainer's outcome measures

We calculated three outcome measures based on the trainers' assessments.

2.3.2.1. Weighted mean overall ability (WMOA). Although trainers scored overall ability at weeks 1, 4, 7, and 10, the ability of the dog at week 10 is more crucial to its future, and having spent 10 weeks training the dog, the trainer's impression is probably more accurate at this stage. Thus, instead of simply averaging all four scores, we weighted the scores according to the number of weeks of training completed.

Weighted mean overall ability = $(1 \times \text{mean overall ability as rated at week 1 by two trainers}) + (4 \times \text{mean overall ability as rated at week 4 by two trainers}) + (7 \times \text{mean overall ability as rated at week 7 by two trainers}) + (10 \times \text{mean overall ability as rated at week 10 by two trainers})/22$.

2.3.2.2. Weighted mean discrepancy from ideal (WMDI). Each trainer ranked each dog for 12 characteristics on four different occasions (Table 1), however, not all characteristics are desirable at high levels; some, such as level of aggression towards humans, may be needed at low levels. During week 10, trainers rated the ideal level of each characteristic that they thought was best suited to specialist search work and we calculated the discrepancy between the rating which each trainer attributed to an individual dog for a give characteristic, and that which they deemed ideal

$$\text{discrepancy} = \sqrt{(\text{ideal score} - \text{individual dog's score})^2}$$

These discrepancy measures were calculated for each trainer, for each of the 12 characteristics, for each dog, and on each of the four assessment weeks. We then calculated the average discrepancy for all characteristics on each assessment week, taking both trainers into consideration. Finally we calculated an overall mean discrepancy for all four assessments, again weighted according to its closeness to the end of training.

Weighted mean discrepancy from ideal = $(1 \times \text{mean absolute discrepancy of 12 traits rated by two trainers at week 1}) + (4 \times \text{mean absolute discrepancy of 12 traits rated by two trainers at week 4}) + (7 \times \text{mean}$

absolute discrepancy of 12 traits rated by two trainers at week 7) + (10 × mean absolute discrepancy of 12 traits rated by two trainers at week 10)/22.

2.3.2.3. Overall ranking (OR). We calculated the mean of the two trainers' ratings for each dogs' future fate (mean future fate). In combination with the scores described above (WMOA and WMDI) this score was used to produce a ranking of all 26 dogs that completed training. Rankings were first made on the basis of mean future fate ratings. Then, amongst dogs with identical scores for this variable, ranks were decided on the basis of their weighted mean overall ability. If dogs tied on the basis of both mean future fate rating and weighted mean overall ability, then weighted mean discrepancy from ideal was used to produce the final ranking.

2.4. Standard search assessment

The assessment involved the dog performing a search, similar to that which would occur during normal operational procedures or during regular training. However we aimed to standardise the handlers' input to allow the behaviour of each of the dogs to be compared directly.

2.4.1. Procedure

The standardised assessment took place during the 10th week of training, in a building used daily for administration, but previously never for dog training or for storage of explosives. The assessments all took place between 1730 and 2030 h, after all staff had left the building.

Five rooms were used and, in four of these, a 75 g sample of one of four different explosives was hidden. Each sample was placed in a standardised hide, which remained the same for all dogs. These included hides at a variety of heights from 0.46 to 1.12 m. To standardise odour dissemination, samples were placed in the hide 1 h before the first dog was tested. Samples were always handled by a person wearing gloves, which were changed between hides to prevent cross contamination.

Each dog was handled by its own trainer (referred to as the handler), and within the group of dogs being assessed on a single day, the order was arranged such that the handlers alternated. Only the two trainers and a camera-operator (on 25 occasions NJR and on one occasion SAG) were present in the building. The procedure for each of the five rooms was the same, and was composed of two phases (a and b).

2.4.1.1. Free search. The handler held the dog by the harness immediately outside the room. The camera-operator entered the room and started filming using a Sony Handycam Vision CCD TRV78E. The second trainer ran into the room, verbally encouraging and stimulating the dog by bouncing a tennis ball. The tennis ball was then hidden on the trainer's person and the handler entered the room with their dog. The handler gave the command "seek on" and then remained close to the door and silent. The dog was ignored for 1 min unless it returned to the handler, or performed a false indication (sitting, freezing or staring in the incorrect direction or more than 2 m from the hide) whereupon the command was repeated.

2.4.1.2. Systematic search. After 1 min, if the dog had not located the hide, the camera-operator informed the handler, who commenced the systematic phase. The handler moved clockwise around the room, using verbal and hand signals to attract the dog to all the prominent features, whilst encouraging the dog to work ahead and independently as much as possible. If the dog failed to locate the hide on the first time around the room, the handler continued on a second circuit. This continued for a maximum of 2 min in the larger rooms 3 and 5 and a maximum of 3 min in rooms 1, 2 and 4. If the dog had not found the hide by the maximum time, it was guided to the correct area and attracted to the hide. Throughout the search, the second trainer moved around the room shadowing the handler, ready to reward the dog. The camera-operator stood near the centre of the room filming the dogs' behaviour.

If the dog sat, or froze and stared, within 2 m of the hide, and orientated in its general direction, and remained there for 5 s, the tennis ball was thrown at the hide. The handler then rewarded the dog by playing with it with the ball for 30 s.

From the video-tapes, two forms of analysis were undertaken:

2.4.2. Subjective ratings

Subjective measures were made by (a) three experienced military dog trainers, none of whom had been involved in the training of these dogs nor were familiar with their performances in training, but all of whom had at least 5 years experience of training specialist search dogs, (“independent trainers”) and (b) three scientists all of whom had worked with dogs and had observed many search dogs during both operational work and training. All observers were trained in a standard way, shown the same sample video and the location of each of the hidden explosives was pointed out. Each observer watched video recordings of each of the 26 dogs during their standard search assessment. They each watched all the dogs over a 3 day period, observing either eight or nine dogs per day. The order in which the dogs were watched was randomised and balanced within the independent trainer and scientist groups. They were requested to watch the videos whilst alone.

Following the completion of an individual dog’s whole test (all five rooms) observers filled in an assessment sheet; on which they rated the dog for each of 11 criteria. These criteria were derived from discussion with expert military dog trainers. They were all characteristics believed to be important in a fully trained search dog and observable from a single assessment period (Table 2). They were each rated on a scale of 1, extremely poor; 2, poor; 3, intermediate; 4, good; 5, excellent. From each observer’s ratings, we compiled a 12th scale of overall total score, derived by summing the scores for each of the 11 attributes.

Table 2

Concordance between raters in their assessment of characteristics measured from the video recordings of the standard search assessment

| Characteristics | Kendall <i>W</i> coefficient of concordance | | | |
|--|---|--------------------------|------------------------|---|
| | Between all six raters | Between three scientists | Between three trainers | Comparing mean of three trainers and three scientists |
| Overall total score | 0.78 ^{***} | 0.91 ^{***} | 0.81 ^{***} | 0.90 ^{**} |
| Motivation/enthusiasm to search | 0.70 ^{***} | 0.84 ^{***} | 0.77 ^{***} | 0.85 [*] |
| General scent recognition—ability to detect explosives’ odours | 0.73 ^{***} | 0.77 ^{***} | 0.82 ^{***} | 0.89 ^{**} |
| Ability to track the scent to source | 0.75 ^{***} | 0.79 ^{***} | 0.88 ^{***} | 0.88 ^{**} |
| Ability to avoid distraction/concentrate on the task | 0.63 ^{***} | 0.78 ^{***} | 0.74 ^{***} | 0.79 [*] |
| Independence—ability to work away from handler | 0.63 ^{***} | 0.75 ^{***} | 0.72 ^{***} | 0.87 [*] |
| Boldness/confidence in new environments | 0.52 ^{***} | 0.45 ns | 0.65 [*] | 0.88 ^{**} |
| Clarity/strength of indication | 0.61 ^{***} | 0.73 ^{***} | 0.66 ^{**} | 0.85 [*] |
| Coverage of area during free search | 0.65 ^{***} | 0.79 ^{***} | 0.70 ^{***} | 0.83 [*] |
| Ability to work without false indications | 0.70 ^{***} | 0.85 ^{***} | 0.79 ^{***} | 0.83 [*] |
| Responsiveness to handler’s command/controllability during systematic search | 0.49 ^{***} | 0.64 ^{**} | 0.57 [*] | 0.75 ns |
| Stamina/fitness to complete the search | 0.54 ^{***} | 0.77 ^{***} | 0.62 ^{**} | 0.72 ns |

The bottom 11 characteristics were each rated on a scale of 1, extremely poor; 2, poor; 3, intermediate; 4, good; 5, excellent. From each observer’s ratings, we compiled a 12th scale of overall total score, derived by summing the scores for each of the 11 attributes.

^{*} $p < 0.05$.

^{**} $p < 0.01$.

^{***} $p < 0.001$.

2.4.3. Objective measures

All objective data recording was done by one trained ethologist (NJR). Each room used for the assessment was mapped in detail, with the separate wall sections identified and all prominent features such as electrical sockets, items of furniture and fire extinguishers outlined. The rooms each had from 10 to 22

Table 3
Objective variables measured from the video recordings of the standard search assessment

| Variable | Description | Transformation |
|--|--|----------------------|
| Free search | | |
| Mean proportion of free search time exploring ^a | Average proportion of one minute free search in each of the five rooms, which the dog spent exploring its surroundings | $\sin^{-1} \sqrt{x}$ |
| Mean duration with people | Average time spent with handler, other trainer, or camera person during free search (maximum of 1 min) | \sqrt{x} |
| Mean proportion of perimeter points covered | Average proportion of the total number of wall sections that the dog contacted or explored within 10 cm of, in each of the five rooms | $\sin^{-1} \sqrt{x}$ |
| Mean proportion of prominent points covered | Average proportion of the total number of searchable objects of wall sections which the dog contacted or explored within 10 cm of, in each of the five rooms | $\sin^{-1} \sqrt{x}$ |
| Mean furthest distance | Average distance from the handler which the dog reached during the free search of each of the five rooms | None |
| Systematic search | | |
| Mean proportion of systematic search time exploring | Average proportion of the five systematic searches that dog spent exploring room | $\sin^{-1} \sqrt{x}$ |
| Mean proportion independent | Average proportion of the five systematic searches that dog spent ahead of, or orientated away from the handler | $\sin^{-1} \sqrt{x}$ |
| Overall | | |
| Mean location scale | Average handler intervention required for dog to pinpoint hide in each of four rooms. Scale: 5, no intervention, located during free search; 4, independently during systematic search; 3, when guided to, during systematic, 2, after precise location indicated by handler, during first circuit of systematic search; 1, after precise location indicated by handler during second circuit of systematic search; 0, not found | None |
| Mean latency to locate hide | Average time taken to locate hide and indicate its position by sitting or freezing and staring, within two metres of, and orientated in the general direction of, the hide(s) | $\log_{10}(x + 1)$ |
| Mean indication latency | Average time between dog showing first interest in the general direction of the hide and clearly indicating its position by sitting or freezing and staring, within two metres of, and orientated in the general direction of, the hide(s) | $\log_{10}(x + 1)$ |
| Mean number of false indications | Average number of times that dog produced conditioned response of sitting or freezing and staring, whilst more than 2 m away from hide, in each of the five rooms | \sqrt{x} |

^a Throughout this paper we refer to the dog exploring, not searching. It is impossible to determine whether the dog is truly “searching”, hence we adopt an objective behaviour terminology of “exploring”.

perimeter points and from 17 to 37 prominent points mapped. This allowed the dog's free search to be monitored closely and measures describing the thoroughness of its search taken. From the videotaped searches, the observer measured 12 objective variables that described the dogs' behaviour during the free and systematic phases of searching (Table 3). The distribution of each variable was examined and transformations employed to improve normality (Table 3).

2.5. Analysis

2.5.1. Trainers' assessments

We examined inter-trainer reliability in the ratings for overall ability and future fate given to individual dogs. The ratings which trainers made following the 10th week of training are likely to be the most accurate since they are based on the most information, so we compared the trainers within each of the two pairs by Spearman's rank correlation tests.

We compared the three derived outcome measures (WMOA, WMDI and OR) using Spearman rank correlation tests.

2.5.2. Standard search assessment—subjective ratings

We compared the ratings for each attribute within the scientists and the independent trainers separately using Kendall's test for concordance (Kraemer, 1979). In addition, we compared the average scores given by these trainers to those given by the scientists (Table 2).

The mean scores for each characteristic (excluding overall total score) were subjected to principal components analysis, to reduce them to the underlying constructs which they represented.

2.5.3. Standard search assessment—objective measures

Using the transformed objective variables, we performed principal components analysis with Varimax rotation to improve the alignment of the measured variables within the factors.

2.5.4. Comparing subjective and objective recording methods

We next compared the outcome measures from the subjective assessment of the 10-week test to the measured objective factors using Spearman's rank correlations.

2.5.5. Comparing trainers' outcome measures and standard search assessment measures

The outcome measures from the trainers' reports were tested for correlation to the subjective and objective measures from the 10-week assessment.

3. Results

3.1. Trainers' assessments

For future fate, seven dogs were rated as suitable for rejection by both trainers (average rating = 1), and 13 as definitely ready to progress to specialist training (average = 4). The other seven dogs required at least some extra training and their ratings were (average, numbers of dogs): (1.5, 1), (2, 2), (3, 3), (3.5, 1).

The future fate ratings given by trainer pair 1 were highly correlated ($\rho = 1.0$) as were those by trainer pair 2 ($\rho = 0.86$) giving an average correlation of $\rho = 0.93$, ($p < 0.001$). Similarly for the trainer's overall ability ratings after 10 weeks of training there was high correlation between the two trainers in pair 1 ($\rho = 0.93$) and pair 2 (0.84) giving an average correlation of $\rho = 0.88$, ($p < 0.001$). Since the trainers within each pair could discuss the performance of their dogs

during training, these high levels of agreement are not unexpected, but justify the subsequent use of within-pair means for trainer ratings.

The three derived outcome measures (WMOA, WMDI and OR) were all highly correlated. We therefore discarded weighted mean overall ability, which was highly correlated to each of the other measures (OR: $\rho = 0.99$, $p < 0.001$; and WMDI: $\rho = -0.84$, $p < 0.001$). We retained the two measures with the lowest correlation; overall ranking and weighted mean discrepancy from ideal ($\rho = -0.82$, $p < 0.001$) and these were used as representative measures of trainers' ratings in subsequent analysis.

3.2. Standard search assessment—subjective ratings

For the majority of the traits, the ratings produced by the scientists and those produced by the independent trainers showed high levels of concordance, higher than either of the groups alone (Table 2). This suggests that both groups of people are rating these traits similarly. Since inter-judge concordance was high for each of the 11 characteristics it was legitimate to use mean ratings for all six judges in subsequent analysis.

Principal components analysis of the 11 subjective variables produced two meaningful factors, each with Eigenvalues greater than 1, and jointly accounting for 83% of the variation in the data. Factor 1 included all but one of the characteristics and is therefore likely to represent “general search ability” (Table 4). Ability to work without false indications was the only variable strongly loaded on factor 2 (loading = 0.96), which therefore represents a separate dimension of working ability.

3.3. Standard search assessment—objective variables

Principal components analysis of the 12 objective variables produced three meaningful factors, factor 1: free search thoroughness (Table 5), factor 2: location ability (Table 6) and factor 3: systematic search behaviour (Table 7), which were therefore retained for subsequent analysis. There was a significant but moderate correlation between factor 1: free search thoroughness and factor 2: location ability ($\rho = 0.51$, $p = 0.008$).

Two of the 12 variables did not appear in any of the three factors. *Mean number of false indications*, although uncorrelated to the other variables, may be an important aspect of search behaviour so was retained as a single variable (factor 4) for future analysis. *Mean indication*

Table 4

Subjective factor 1: general search ability—key variables that combine to produce this factor and their respective loadings

| Variable | Loading |
|--|---------|
| Positively loaded variables | |
| Ability to avoid distraction/concentrate on the task | 0.96 |
| General scent recognition—ability to detect explosives' odours | 0.94 |
| Motivation/enthusiasm to search | 0.94 |
| Stamina/fitness to complete the search | 0.92 |
| Independence—ability to work away from handler | 0.91 |
| Ability to track the scent to source | 0.91 |
| Coverage of area during free search | 0.90 |
| Clarity/strength of indication | 0.88 |
| Responsiveness to handler's command/controllability during systematic search | 0.79 |
| Boldness/confidence in new environments | 0.72 |

Table 5

Objective factor 1: free search thoroughness—key variables that combine to produce this factor, and their respective loadings

| Variable | Loading |
|---|---------|
| Positively loaded variables | |
| Mean furthest distance | 0.89 |
| Mean proportion of perimeter points covered | 0.83 |
| Mean proportion of free search time exploring | 0.71 |
| Mean proportion of prominent points covered | 0.68 |
| Negatively loaded variables | |
| Mean duration with people | 0.88 |

Table 6

Objective factor 2: location ability—key variables that combine to produce this factor, and their respective loadings

| Variable | Loading |
|-----------------------------|---------|
| Positively loaded variables | |
| Mean location scale | 0.90 |
| Negatively loaded variables | |
| Mean latency to locate hide | 0.85 |

latency was relatively difficult to measure and, since it did not align with any of the factors, this variable was discarded.

3.4. Comparing subjective and objective recording methods

Subjective factor 1: general search ability was significantly associated with both objective factor 1: free search thoroughness ($\rho = 0.66$, $p < 0.001$) and objective factor 2, location ability ($\rho = 0.82$, $p < 0.001$). The subjective factor 2, ability to work without false indications was significantly negatively correlated to the objective factor 4, mean number of false indications ($\rho = -0.87$, $p < 0.001$).

3.5. Comparing trainers' outcome measures and standard search assessment measures

Both of the trainers' outcome measures showed significant correlation to the subjective factor 1: general search ability (overall ranking: $\rho = 0.74$, $p < 0.001$; weighted mean discrepancy from ideal: $\rho = -0.75$, $p < 0.001$; Table 8). In contrast, no trainers' outcome measures showed significant association to the subjective factor 2, ability to work without false indication.

Table 7

Objective factor 3: systematic search behaviour—key variables that combine to produce this factor, and their respective loadings

| Variable | Loading |
|---|---------|
| Positively loaded variables | |
| Mean proportion of systematic search time exploring | 0.93 |
| Mean proportion independent | 0.66 |

Table 8

Correlations between trainers' assessments and measures taken during the standard search assessment, both subjective and objective

| Standard search assessment measure | Trainers' assessment outcome measures | |
|---|---------------------------------------|--------------------------------------|
| | Overall ranking | Weighted mean discrepancy from ideal |
| Subjective measures | | |
| Factor 1: general search ability | 0.74*** | −0.75*** |
| Factor 2: ability to work without false indications | −0.18 ns | 0.06 ns |
| Objective measures | | |
| Factor 1: free search thoroughness | 0.60*** | −0.48* |
| Factor 2: location ability | 0.54** | −0.62*** |
| Factor 3: systematic search behaviour | 0.47* | −0.44* |
| Factor 4: mean number of false indications | 0.16 ns | −0.09 ns |

ns: non-significant.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

Objective factors 1–3 from the 10-week assessment also correlated significantly with both of the trainers' outcome measures (Table 8). The objective factor 4, mean number of false indications, showed no significant association to either of the trainer's outcome measures ($\rho < 0.17$, $p > 0.4$).

The objective measures generally showed lower correlations than the subjective ones. However, since there are several objective factors all describing different aspects of search behaviour, it is possible that the trainers combined several of these factors to derive their assessments. To determine whether this is true, we used canonical discriminant function analysis (Manly, 1986). We divided the dogs into four groups according to their mean future fate rating. We then performed discriminant analysis to examine what combination of objective factors best predicted the dogs' groupings. This analysis produced one significant dimension ($\chi = 37.7$, $p < 0.001$), the main contributors to which were factor 1 (correlation coefficient = 0.62) and factor 2, (0.55) followed by factor 3 (0.38), whilst factor 4 contributed very little (0.04). This confirms that the objective factors, in combination but excluding false indications, do reliably predict a dog's outcome in training.

4. Discussion

This is the first time, to our knowledge, that an empirical method for assessing the ability of search dogs has been validated and documented. We have quantified the ability of 26 dogs via measures derived from their trainers' ratings and via subjective and objective measures of their performance in a standardised search assessment. There was a high level of agreement between these methods.

The standard search assessment was used to give a single point sample of the dogs' behaviour and ability. This proved to be a relatively easy assessment procedure to perform in a standardised way, and from the video recordings of the assessment, two types of analysis were undertaken. Six independent dog trainers and scientists rated the dogs on 11 aspects of performance, and there was generally a high level of agreement between the observers. There was significant concordance both within the groups of scientists and independent trainers, and between all six

observers. This result shows that the 11 behavioural traits can be consistently judged and that all observers possessed a similar idea of what each trait constituted. These characteristics are therefore potentially useful when dogs' search ability is being rated by multiple observers. The fact that scientists, who had extensive experience of observing search work, rated the dogs similarly to those who had themselves trained dogs, confirms that it is legitimate to utilise trained scientists to provide subjective data on proficiency in future trials and studies of search dogs. As long as the scientists are very familiar with the search task in question, their opinions are unlikely to differ significantly from skilled search dog operatives.

When the subjective ratings made by the observers were subjected to PCA, ratings for many of the 11 characteristics fell into two distinct groups; one describing a general ability to search, and the other, a tendency to provide incorrect (false) indications. This suggests that although there was consistency in the observers' ratings for all 11 traits, their scores were in fact describing only two underlying important traits.

When the same search task was analysed objectively, the measures of the dogs' behaviour produced four factors; free search thoroughness, location ability, systematic search ability and mean number of false indications. Although given our sample size (26) we cannot be confident that the factors we have identified would be adequate to describe all search scenarios or other populations of dogs. The first two of these appeared to correspond to aspects of the subjective ratings of general search ability, whilst the fourth corresponded to the subjective rating of ability to work without false indications. This shows that the subjective ratings, which the observers made were accurate and were based on behaviours which are quantifiable. The subjective and objective factors describing dogs' tendency to give false indications correlated highly, hence each method of analysis produces a good representation of this behavioural trait. In the case of general search ability, the objective measures produced two factors, which in combination correlated to this subjective factor. Thus, by using objective empirical measures we were able to dissect search performance into a greater number of dimensions, compared to the use of subjective ratings. This may be beneficial, as by concentrating on specific aspects of performance it may be possible to enhance the ability of future search dogs by refining training, changing procurement methods, directing remedial training or even breeding dogs specifically for a given behavioural dimension. Thus, objective measures do have considerable value in allowing individual components of ability to be considered in isolation. However subjective ratings are considerably less time-consuming, and give a good overall representation of ability. Since both types of measures are highly correlated we conclude that they are both give reliable information, and that either one could be used to assess dogs, depending upon the level of detail required.

The subjective and objective ratings were taken during a single testing session, and in order to be useful they must be representative of the dog's usual performance. When comparing the assessment measures to the trainers' ratings of the dogs' ability, the majority of the measures correlated significantly. The subjective rating for general search ability was extremely highly correlated to trainer's ratings. Of the four objective measures, free search thoroughness and location ability combined with systematic search behaviour, all correlated strongly with the trainers' rankings of the same dogs. However the trainers completed the final ratings of the dogs' ability before handling some of them during the standard search task. There was therefore a possibility that the trainers' opinions of the dogs' ability influenced their behaviour (consciously or subconsciously) towards the dog, which in turn affected performance, a Clever Hans effect (Martin and Bateson, 1993). With some working dog roles this possibility could have been overcome by using naïve handlers, however in the case of explosives search dogs, the nature of the training and operation means that dog and handler, or dog and trainer, must have formed a

relationship before they can work effectively as a team. In this case we can assess the likelihood of such an effect by considering the free search phase of the standard search task, during which the trainer stood still and had minimal interaction with the dog. If a Clever Hans effect were apparent, one would expect the trainers' assessments to show weaker correlations to the dogs' performance during this free search than to the dogs' behaviour during the closely guide systematic search. In fact the converse is true; the objective factor free search thoroughness showed a stronger correlation to the trainers' ratings than did the factor systematic search behaviour (Table 8). We believe this is evidence against a significant Clever Hans effect, and conclude that standard search task method is therefore a reliable indicator of the success of training, and a useful tool for comparing individual dogs in a standard way.

Measures of the dogs' tendency to false indicate during the standard search assessment were not correlated to their trainer's ratings of ability. Therefore, either false indications in this assessment were not representative of the dog's usual behaviour, or, more likely, trainers do not consider a tendency to false indicate an important criterion when judging their dogs. However, whilst a dog giving a false positive indication of the position of an explosive in a training or testing environment may not be a great problem, in an operational environment it may constitute a much greater concern. A false indication may cause a handler to declare a high alert search, with considerable time and financial cost. Thus, we would suggest that perhaps a tendency to give false indications is a behavioural characteristic, which may require additional attention during the selection and training of search dogs.

Although subjective and objective ratings of behaviour are both used in studies of dog behaviour, there have been very few previous studies comparing their respective outcomes, or examining the validity of any of the measures. Studies involving subjective assessment of dogs' performance in guide dog suitability tests have shown low inter-observer reliability (Goddard and Beilharz, 1982; Murphy, 1995). However we saw relatively high inter-observer correlations in our measures. We suggest that this may be because of the detailed process by which we developed the subjective characteristics to be rated (Rooney and Bradshaw, 2004; Rooney et al., 2004) and the fact that we bench-marked our scales with specific descriptors. In contrast, when comparing owners' subjective ratings of several behavioural traits, Rooney (1999) saw limited correlations to measures of the dogs' actual behaviour. We suggest this difference arises because, in the current study, we were very clear in the description of the traits which the observers were required to rate; all terminology was defined and previously tested for clarity in a similar population. In addition, unlike the dog owners, the observers in this study all rated numerous animals and hence they could make comparisons, which led their judgements to be more internally reliable. Similarly when looking at another species, Wemelsfelder et al. (2001) found that subjective descriptions of pig behaviour were consistent between observers, suggesting that humans can reliably rate the behaviour of other domestic animals.

5. Conclusion

The standard assessment of search dog ability is a convenient, economical and effective way of assessing a dog's ability at search work. It corresponds well to trainers' ratings of dogs' ability, but also adds information. Therefore this assessment technique is potentially useful for comparing the relative effectiveness of different dogs, different breeds, different training methods and different agencies. It may thus be useful for future research, and similar methods of assessing performance may prove applicable to many working dog roles. We have demonstrated how both objective and subjective assessment of behaviour have value in the assessment of behaviour.

Acknowledgements

We thank all the staff at the Defence Animal Centre who helped with this study. Particularly thanks to Colonel Macdonald, Warrant Officer Brown and Major Ham, who offered endless help and support. Thank you to the DAC trainers, Staff Sergeant Alun Hodges, Sergeant David Whelton, and Corporal Amanda Swanick and to scientists Darren Saunders, David Lewrey and Sara Jackson for completing the inter-observer assessments. We are also grateful to Flight Sergeants Kenny Braddick and Dave Blundell for helping us devise the assessment and training methodologies. Finally our sincere thanks to our trainers: Corporal Tracey Penman, Corporal Mark Cope and Lance Corporal Emma Davies. They were a joy to work with and showed a commitment to and enthusiasm for our study that went far beyond the call of duty.

References

- Gazit, I., Terkel, J., 2003. Explosives detection by sniffer dogs following strenuous physical activity. *Appl. Anim. Behav. Sci.* 81, 149–161.
- Goddard, M.E., Beilharz, R.G., 1982. Genetics of traits which determine the suitability of dogs as guide dogs for the blind. *Appl. Anim. Ethol.* 9, 299–315.
- Kraemer, H.C., 1979. Ramifications of a population model for K as a coefficient of reliability. *Psychometrika* 44, 461–472.
- Manly, F.J., 1986. *Multivariate Statistical Methods: A Primer*. Chapman and Hall, London, UK.
- Martin, P., Bateson, P., 1993. *Measuring Behaviour: An Introductory Guide*, 2nd ed. Cambridge University Press, Cambridge, p. 33.
- Murphy, J.A., 1995. Assessment of the temperament of potential guide dogs. *Anthrozoos* 8 (4), 224–228.
- Rooney, N.J., 1999. Play behaviour of the domestic dog (*Canis familiaris*) and its effects on the dog–human relationship. PhD thesis, University of Southampton.
- Rooney, N.J., Bradshaw, J.W.S., 2004. Breed and sex differences in the behavioural attributes of specialist search dogs—a questionnaire survey of trainers and handlers. *Appl. Anim. Behav. Sci.* 86, 123–135.
- Rooney, N.J., Bradshaw, J.W.S., Almey, H., 2004. Attributes of specialist search dogs—a questionnaire survey of UK dog handlers and trainers. *J. Forensic Sci.* 49, 300–306.
- Schoon, G.A.A., 1996. A first assessment of the reliability of an improved scent identification line up. *J. Forensic Sci.* 43, 70–75.
- Schoon, G.A.A., 1998. Scent identification lineups by dogs (*Canis familiaris*): experimental design and forensic application. *Appl. Anim. Behav. Sci.* 49, 257–267.
- Schoon, G.A.A., De Bruin, J.C., 1994. The ability of dogs to recognise and cross match human odours. *Forensic Sci. Int.* 69, 111–118.
- Serpell, J.A., 1996. Evidence for an association between pet behaviour and owner attachment levels. *Appl. Anim. Behav. Sci.* 47, 49–60.
- Shivik, J.A., 2002. Odor-absorptive clothing, environmental factors, and search-dog ability. *Wildl. Soc. Bull.* 30, 721–727.
- Svartberg, K., Tapper, I., Temrin, H., Radesäter, T., Thorman, S., 2005. Consistency of personality traits in dogs. *Anim. Behav.* 69 (2), 283–291.
- Wemelsfelder, F., Hunter, T.E.A., Mendl, M.T., Lawrence, A.B., 2001. Assessing the 'whole animal': a free choice profiling approach. *Anim. Behav.* 62, 209–220.
- Williams, M., Johnston, J.M., 2002. Training and maintaining the performance of dogs (*Canis familiaris*) on an increasing number of odor discriminations in a controlled setting. *Appl. Anim. Behav. Sci.* 78, 55–65.
- Wilsson, E., Sundgren, P.E., 1997a. The use of a behaviour test for selection of dogs for service and breeding. 1. Method of testing and evaluating test results in the adult dog, demands on different kinds of service dogs, sex and breed differences. *Appl. Anim. Behav. Sci.* 53, 279–295.
- Wilsson, E., Sundgren, P.E., 1997b. The use of a behaviour test for selection of dogs for service and breeding. 2. Heritability for tested parameters and effect of selection based on service dog characteristics. *Appl. Anim. Behav. Sci.* 54, 235–241.