Applied Animal Behaviour Science 165 (2015) 133–142

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim

Validation of a short odour discrimination test for working dogs☆

Fay Porritt*, Ralph Mansson, Alison Berry, Natalie Cook, Nicola Sibbald, Steve Nicklin

Dstl, Fort Halstead, Sevenoaks, Kent TN14 7BP, United Kingdom

ARTICLE INFO

Article history:
Accepted 30 November 2014
Available online 9 December 2014

Keywords:
Odour
Discrimination
Dog
Detection
Explosive
Validation

ABSTRACT

A short odour discrimination test has been designed to allow rapid quality assurance of odour recognition by detection dogs. The test comprises five repeats per target and a minimum of 20 associated non-target odours. The mean time taken to conduct the test is 5.6 min per target type. A pass criterion of "a detection rate at least 70% greater than false alarm (FA) rate, with a 15% cap on total allowable false alarms" is used which equates to 4/5 correct indications and 2 FAs, or 5/5 correct indications and 3 FAs; the probability of passing this test by chance is <1%. A Microsoft Excel™ programme has been written to rapidly generate balanced running orders that allow search runs to be truncated following correct indications; this speeds up testing whilst maintaining standardisation; the programme is available free-to-use. The test's internal validity has been measured by conducting test re-test analysis on a range of target types on 19 operational search dogs, and external validity has been measured by completing the test and an equivalent operationally relevant building search on 26 operational search dogs. In both cases there is good overall reliability (kappa ≥ 0.80). The test is thus deemed suitable for complementary assessment of detection dog ability during detailed accreditation procedures or as a standalone quality assurance test in between accreditation or licensing.

1. Introduction

Dogs have been used in scent detection tasks for many years and are currently widely deployed for the detection of contraband such as drugs and explosives across the world (Gazit and Terkel, 2003; Bird, 1997). In recent years the use of scent detection dogs has expanded rapidly with such dogs now being used for a wide range of conservation detection tasks and in an increasing number of medical screening research programmes (see Johnen et al., 2013).

In the majority of these roles, the initial confirmation of a dog's ability is determined using some form of odour discrimination task to provide a measure of accuracy (proportion of targets detected) and specificity (proportion of non-targets correctly discriminated against); ongoing quality assurance tests of operational dogs are rarely reported in the peer reviewed literature. As the statistical reliability of both of these test measures is based on the number of samples included in the test (target and

☆ Dstl© Crown copyright (2014). Licensed under the terms of the Open Government Licence except where otherwise stated. To view this licence, visit http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3 or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or psi@nationalarchives.gsi.gov.uk.

* Corresponding author at: Fort Halstead, S18 rm41, Sevenoaks, Kent, TN14 7BP, United Kingdom. Tel.: +44 01959 892385; fax: +44 01959 892506.
E-mail address: fporritt@dstl.gov.uk (F. Porritt).

http://dx.doi.org/10.1016/j.applanim.2014.11.021

non-target respectively), the tests are by necessity time consuming and labour intensive, with typical tests including 25–50 samples per target and three to six times as many non-target samples (for review see Johnen et al., 2013). As a consequence, tests are typically reported as taking several days to complete (e.g.14 days Richards et al., 2008; 10 days Lin et al., 2011; 2 days Ehmann et al., 2012).

While the length of these tests is acceptable for research dogs or dogs trained to detect individual odours, the majority of dogs employed in contraband detection are trained to detect multiple targets often numbering in the tens or higher (e.g. Williams and Johnston, 2002). In addition, unlike dogs used for remote detection tasks, such as medical screening, where samples are delivered to a central location and presented in a controlled set-up, contraband dogs are also required to maintain a high skill-level in their search technique and ability to search in multiple operational locations.

The maintenance of the high skill level of these dogs requires significant training time alongside operational duties, and as there is a cost associated with any extraction from work, training time is at a premium. Despite this pressure on training time, both initial accreditation and ongoing quality assurance are critical, particularly in roles where lives are at risk if dogs do not perform to their maximum potential, e.g. explosives detection. Any reduction in the time taken to conduct quality assurance or increase in the standardisation of testing would therefore be beneficial.

The different organisations that employ detection dogs have their own in-house accreditation and ongoing quality assurance requirements. The most significant component of these tests is the measurement of the dog and handler teams' ability to conduct safe, systematic searches using appropriate procedures for the organisation in question; this area of testing has been addressed by several authors (e.g. Diederich and Giffroy, 2006; Rooney et al., 2007). However a second component of detection dog performance is the ability to correctly discriminate all trained odours from non-trained odours; this component of performance is often subsumed in the larger test of general ability. As the general test of ability is almost always based on search scenarios, this approach results in a lack of standardisation for odour discrimination (targets will be placed in different locations on all searches), and also has the disadvantage of being very time consuming if multiple presentations of each target are to be carried out.

The aim of this study was to develop and validate a short test for the quality assurance of odour discrimination ability (not search ability) for use in initial accreditation and ongoing quality assurance of working contraband detection dogs. Any test should fit four basic quality requirements (Diederich and Giffroy, 2006; Martin and Bateson, 1986); the administration and notation of the test should be standardised with the only variable being the animal tested; the test must be reliable, if it is conducted twice the results should be significantly correlated; the test should be sensitive enough to give a meaningful measure of performance, using a precise and objective scale, and finally the test must be valid. Validity is split between internal validity, does the test measure what it pertains to measure, and external validity, is performance in the test predictive of performance in a relevant real world task (Diederich and Giffroy, 2006).

This test must also fit additional criteria if it is to be of utility to the widest range of practitioners; it should be suitable for use by non-scientists with minimal training, standardisation must be achieved in a non-laboratory setting without direct input from scientists or test originators, and finally, the test must place the minimum logistic burden on those conducting it both in material costs and time to conduct. This final requirement will by necessity result in a trade-off between test sensitivity/validity and logistic burden. The aim was therefore to devise a test which reliably differentiated between at least two groups of dogs (acceptable and unacceptable).

A test meeting these requirements would be suitable for use in ongoing quality assurance of odour discrimination ability of working dogs where extended periods of extraction from work for extensive quality assurance are not feasible; such a test would also complement initial accreditation of dogs if used alongside other more extensive and time consuming testing methods.

2. Methods

2.1. Subjects

All subjects were police explosives detection dogs (EDDs) or UK Border Force drug detection dogs (DDDs). All dogs had been trained by police or border force officers to search for a range of explosives or drugs following standard Association of Chief Police Officer (ACPO) training guidelines which rely on shaping required behaviours using positive reinforcement (play). All dogs had successfully passed initial ACPO licensing and additional re-certifications where appropriate, confirming that that they were able to search safely and were able to locate and indicate on all required targets for their role. All dogs had completed at least one month of operational duties (range 1 month–8 years, mean 4.5 years).

Thirteen dogs (7 male and 6 female) were used during pilot testing and 45 dogs (27 male and 17 female) were used during validation; breeds included Labradors, English and working-cocker spaniels, English and German pointers and cross breeds. Prior to training conducted for this study, subjects had not received training on odour discrimination line-ups; however they were all able to discriminate between target and non-target odours whilst conducting searches.

2.2. Procedure outline

The test was a multi choice experiment, equivalent to other odour discrimination procedures used by dog practitioners and research scientists in multiple agencies around the world (see Johnen et al., 2013). This set up requires a handler and dog to walk down a line of numbered identical stainless steel sample containers which contain either an "interferent" odour, that the dog must ignore or a "target" odour, that the dog should indicate on by sitting or freezing whilst orientated to the target sample (Fig. 1). Samples were placed 1 m apart in wooden (pilot testing) or Perspex

Fig. 1. Odour ID test set up showing first four stands and removable stainless steel tins.

holders (validation), referred to as stands. The 1 m distance allowed large targets of up to 5 g to be used without any apparent interference between adjacent samples. Each line of samples is referred to as a "run"; dogs were required to search the run in order on a lead. There was no time limit for the search but dogs were not permitted to go back to a stand once they had moved on to the next one; dogs therefore searched each sample once.

To minimise running time all tins containing interferent odours were pooled together during testing and two tins were chosen at random to switch into the line after each run. Any samples that were falsely indicated on were removed and replaced with new interferents in clean tins. All sample tins and lids were permanently marked with target type or generic "interferent" underneath to reduce possibility of cross contamination between target and non-target tins. All tins were washed in a dishwasher at the start and end of testing.

The test was conducted double blind by one assessor (two during pilot testing) and data were collected based solely on handler declarations. Handlers stated the stand number when they believed their dog had indicated on a target, the researcher, who was screened from view, verbally responded "reward" following a correct indication or "no" following an incorrect indication. If handlers believed that a target was present but their dog did not give a full indication, they were instructed to say "interest" in which case no information was given to the handler and the search continued.

Following an incorrect indication, commonly referred to as a "false alarm" in working dogs, the handler and dog left the search area and all sample tins up to and including the sample indicated on were removed. The handler and dog then returned to the area and searched all empty stands and remaining samples.

During pilot testing dogs were rewarded for a correct indication, samples up to and including the target were removed as per a false alarm and dogs were then required to search the remaining samples in the line. However in the final test procedure dogs were rewarded following a correct indication and the dogs did not search the remaining samples in the run.

2.2.1. Scoring

A separate detection rate was calculated for each target odour, with each correct indication counting as +1 and each correct "interest" counting as +0.5. This approach allows dogs to be given remedial training and re-testing only on odours of concern and it also allows testing to fit around operational commitments by giving freedom to assessors to test on a range of target numbers per session.

A total false alarm score was calculated per dog per test session, with each false alarm counting as +1 and each false interest counting as +0.5. The total false alarm score in the session was divided by the number of target types tested and this average score was used to determine whether the dog had an acceptable false alarm rate. As with the detection rate, this approach allows dogs to be tested around operational commitments, it also ensures that dogs with good detection but a problem with discrimination, receive different remedial training than dogs with good discrimination but poor performance on a specific target.

The pass criteria for detection and false alarms are covered in Section 2.3.2.

2.3. Design

The aim of the design stage was to develop the shortest possible test that retained acceptable type 1 and type 2 errors.

2.3.1. Test precision

The precision of detection rate estimates was determined for a range of short test lengths by calculating exact 95% confidence intervals for four to six presentations of a target for dogs with true detection rates of 40%, 80% and 95% (see Clopper and Pearson, 1934 and R Core Team, 2012).

2.3.2. Pass criteria and test confidence

A range of pass criteria from "Pass 60" to "Pass 95" (P60–P95) were generated based on signal detection theory. In each case the criterion states that the dog passes the test if the proportion of targets found by the dog is a set percentage greater than the proportion of non-targets which a dog false indicates on:

$$\text{Pass},R = \left(\frac{x}{n} - r > \frac{y}{m}\,(\text{pass if detection rate is } R\%\text{ better than false alarm rate}\right)$$

where R = percentage that detection rate must be greater than false alarm rate to pass, x = number of indications on a true target; n = total number of targets in test; $r = R/100$, y = number of indications when there is no target (false alarms); m = total number of blanks in test calculated as (number of runs) × (number of stands) − n.

Each pass criterion gives a different trade-off between the number of poor quality dogs that will pass the test by chance and the number of high quality dogs that will fail the test by chance. These interactions were plotted on contour maps for P60–P95 pass criteria and presented to five canine detection subject matter experts. Following subjective assessment of the trade-offs, a pass criterion of P70 was unanimously agreed.

For P70 criterion, alpha values were calculated for four to eight repeats per target, with 10–25 interferent samples per target, for dogs with 0%, 10%, 20%, 30% and 40% true detection rate. These calculations were used to determine the minimum number of interferent samples required to give an acceptable test confidence.

2.3.3. Odour ID test procedure

Based on test precision and confidence analyses (see Section 3.1), a test containing five repeats per target type and a minimum of 20 interferents per target type was chosen and used for validation stages.

2.4. Targets

No single targets were identified (during pre-trial tests) that gave a full range of outcomes for the target population. Therefore in order to generate the variability required to validate a test, a range of seven target types were used including targets that the subjects were not trained to detect. Targets were graded by anticipated difficulty or detection probability and were named according to their ranking (target 1 = easiest, target 7 = most difficult) (Table 1).

The use of multiple targets to show variation requires an assumption that searches for each target type are independent events. This was tested in an unrelated experiment (in prep.) that found that presentation of six consecutive "Easy" targets or six consecutive "Difficult" targets did not affect the detection rate of "Difficult" targets presented immediately afterwards (general linear model, $P > 0.05$). The assumption of independence of target outcomes was therefore used for data analyses in this experiment.

2.5. Interferents

Twenty five different interferent odours were used for each test session; these were 1–5 g of everyday items including 12 strongly perfumed odours such as shampoo, soap and nylon gloves and 13 substances without a perfumed odour such as soil, cotton wool, dried pasta and plastic bags. In addition five empty sample pots were used as interferents in each test session.

2.6. Pilot testing

Pilot testing investigated the effect of multiple targets in one run and the logistics of the general test set up. All dogs were given two 20 min training sessions with the equipment prior to pilot testing.

Thirteen operational EDDs each completed 19 runs of six stands; three runs contained six interferents and no targets, the remaining runs were made up of variable target samples, with the total number of samples made up to six per run with interferents as follows: four runs contained two target samples (target 1 and target 5), six runs contained one sample of target 1 and six runs contained one sample of target 5. A total of ten samples of each target type were used and all dogs searched the same samples.

All indications, false alarms and "interest" calls given by handlers were recorded by the assessor. In addition a subjective assessment of whether the dogs searched each pot was made by a second researcher. chi^2 test was used to determine whether there was any effect of multiple targets per run. Post hoc chi^2 tests were used to determine whether there was uneven searching of sample pots and uneven distribution of false alarms.

2.7. Odour ID test set up

Based on the outcomes of pilot testing the final protocol included a maximum of one target per run. The number of samples per target type was set at two as this was the minimum that could be used to run the test smoothly; the effectiveness of this approach was addressed through empirical validation stages.

In addition an extra stand was added to the start and the end of each run to give a total of eight stands per run, these stands only ever included interferent odours during testing, and handler were informed of this procedure (targets or interferents were placed in these stands during training). Stands one and eight were not included in the calculation of test power and confidence and any false indications on these stands were excluded from all. These stands act as tools to ensure that stands two to seven were correctly searched. Stand one ensured that dogs were searching prior to encountering the first potential target, and stand eight allowed handlers to ignore false alarms that were given due

Table 1
Targets used for validation experiments for explosives detection dogs (EDDs) and drug detection dogs (DDDs).

Target name	Predicted difficulty (rank)	Identity	Dogs	Expected detection rate
Target 1	Very easy (1)	5 g trained explosive	EDDs	High – dogs trained to detect target
Target 2	Easy (2)	1 g trained drug	DDDs	High – dogs trained to detect target
Target 3	Intermediate (3)	2 g trained explosive	EDDs	Medium/high – trained on similar target in different quantity
Target 4	Intermediate (4)	5 g of untrained explosive	EDDs	Medium – not trained on target but some generalisation expected
Target 5	Difficult (5)	5 g of untrained simulated explosive (non-explosive)	EDDs	Low – not trained on target and low or no generalisation expected
Target 6	Difficult (6)	0.023 g trained explosive	EDDs	Low – subjects selected for this study had not been trained to indicate on this mass
Target 7	Should not indicate (7)	5 g untrained explosive (drug detection dog)	DDDs	None – dogs not trained on similar targets, no generalisation expected

to dogs indicating on the last stand in the "hope" of getting a reward.

2.7.1. Standardisation

Limiting the number of targets to one per run removes the independence of samples searched after a correct indication as the team will be aware that there can be no more targets in the run. This is avoided if the search is ended after each correct indication (truncated); however this reduces standardisation across dogs as some dogs will search significantly fewer interferents than others.

As a solution to this problem, a Microsoft™ Excel programme was written that maintains the standardisation of truncated searches by returning a randomly selected "running order" from a pre-generated set of 80 randomised designs. Each running order ensures that a dog with 100% detection rate searches an average of 4.5 interferent samples per target repeat, with a tolerance of up to five samples (e.g. 4.5 × 5 repeats = 20–25 interferent samples per target type); a dog with a lower detection rate will search a small number of extra samples. These running orders also contain balanced target positions such that half of all targets occur in positions 2–4, and half occur in 5–7.

Using this software, following a correct indication, dogs were rewarded and the run was recorded as complete; actions following false alarms and interest calls were unchanged (see Section 2.2). The alpha value was calculated for the final test protocol (5 repeats per target type plus mean 22.5 interferents per target type) and this protocol was used for all validation experiments and is referred to as "the odour ID test".

2.8. Validation

2.8.1. Internal validity: repeatability

Thirteen EDDs and six DDDs were given two 20 min training sessions on the odour ID test. The EDDs then completed the odour ID test on targets 1, 4 and 5 and the DDDs completed the odour ID test on targets 2 and 7 (see Table 1). All dogs repeated the same test the following day to allow test re-test analysis. Two samples of each target type were used for the first test following the odour ID test protocol; however five different samples of each target type were used for the re-test. This was to confirm that performance on two samples of a target was predictive of performance on five samples of a target, thereby validating the use of only two samples per target type.

2.8.2. External validity: comparison of odour ID and search performance

UK Police EDD units were provided with the odour ID test protocol, software, equipment and training and used the test during their regular training over a 12-month period.

Following this period, 18 randomly selected EDDs completed the odour ID test on targets 3 and 6 (2 samples of each target) and completed a building search containing five different samples of targets 3 and 6 (counterbalanced odour ID or building first). A different eight EDDs followed the same procedure using the targets 1 and 5.

The building search was separated into 12 rooms; each room contained six odours concealed in furniture at ≤1 m height. For each target, five rooms contained the target plus five interferents and one room contained six interferents; each room was therefore equivalent to one run for analysis.

Handlers conducted a systematic building search based on Rooney et al. (2007). This consisted of 20 s free search with no direction from the handler, followed by a directed search in a counter clockwise direction around the room. Handlers verbally reported all indications and interest, and actions following an indication were the same as for the odour ID test.

This component was single blind as the observer was aware of the placement of all interferents and targets but the handler was unaware. To reduce observer influence, observers always stood on a mark on the floor and the same data were recorded for targets and interferents including whether the dog searched the target; this ensured that observers were equally likely to look at interferents and targets.

2.8.3. Analyses

No dogs failed due to excessive false alarms; that is all dogs with 4/5 had two or fewer false alarms and all dogs with 5/5 had three or fewer false alarms (as allowed by the pass criterion), the false alarm rate is therefore excluded from all analyses for clarity.

For each experiment, Cohen's kappa was used to indicate test re-test repeatability of the combined data from all targets.

For both experiments, performance on each test re-test event was scored as pass ≥4/5 or fail <4. Performance was also categorised as one of three groups; poor (fail, <3), medium (fail; 3–3.5) or good (pass; 4–5). Exact 95% confidence intervals were calculated to estimate the likelihood of a dog from each group passing or failing a future odour ID test.

3. Results

3.1. Test precision

Exact 95% confidence intervals were calculated to indicate the precision of the detection rate outcome from a test containing four, five or six repeats per target (Table 2). These data highlight that at least five repeats per target are necessary to correctly fail poor dogs (true detection 40%) and pass very good dogs (true detection 95%) on at least 90% of tests. Six repeats per target is better at correctly failing poor dogs but leads to a lower pass rate for good and very good dogs (Table 2).

3.2. Pass criteria

A pass criterion of P70 was chosen by subject matter experts, however a total cap of 15% was placed on permitted false alarms as 30% was felt to be too high. Using this pass criterion, 20 interferent samples per target type were sufficient to give alpha $P < 0.001$ for dogs with 0% true detection following four, five or six target repeats and $P < 0.05$ for dogs with 10–20% true detection rate (Fig. 2).

Table 2
Exact 95% confidence intervals for four to six repeats per target showing an unacceptably high pass rate (18–35%) for poor quality dogs with four repeats and >90% precision for poor and very good dog categorisation for five and six repeats (*). Table also shows a better precision for good dogs following five rather than six repeats per target (74 vs. 66%).

Repeats per target	Pass criteria (detection %)	True detection (%) – dog ability	Dogs expected to pass (%)
4	3/4 (75)	40 – poor	18
		80 – good	82
		95 – very good	99
5	4/5 (80)	40 – poor	9*
		80 – good	74
		95–very good	98*
6	5/6 (83)	40 – poor	4*
		80 – good	66
		95–very good	97*

For dogs with 30% true detection rate, neither 20 nor 30 interferents were able to maintain alpha $P < 0.05$ for four to six repeats (Fig. 2).

Based on these analyses, a test length of five target repeats with a minimum of 20 interferents and a pass criterion of P70 (capped at 15% false alarm) was used for testing. This equates to an actual detection pass criterion of 80% plus ≤2 false alarms, or 100% plus ≤3 false alarms when dogs are presented with five repeats per target type.

3.3. Pilot testing

The presence of two targets in one line did not significantly affect the detection rate on the first or second target (chi^2, $P > 0.05$). However runs with two targets were logistically difficult to set up and resulted in errors by assessors (two targets left out after runs, one incorrect target used).

Dogs were significantly less likely to search the first stand compared to all other stands (13% vs. 2.5% respectively; chi^2, $P > 0.05$). Likewise, dogs were significantly more likely to false indicate on the last stand compared to all other stands (7% vs. 1.7% respectively; chi^2, $P > 0.05$).

3.4. Test alterations: standardisaton

Based on the outcomes of pilot testing, the test was altered to include a maximum of one target per run; and a Microsoft Excel™ programme was written to standardise

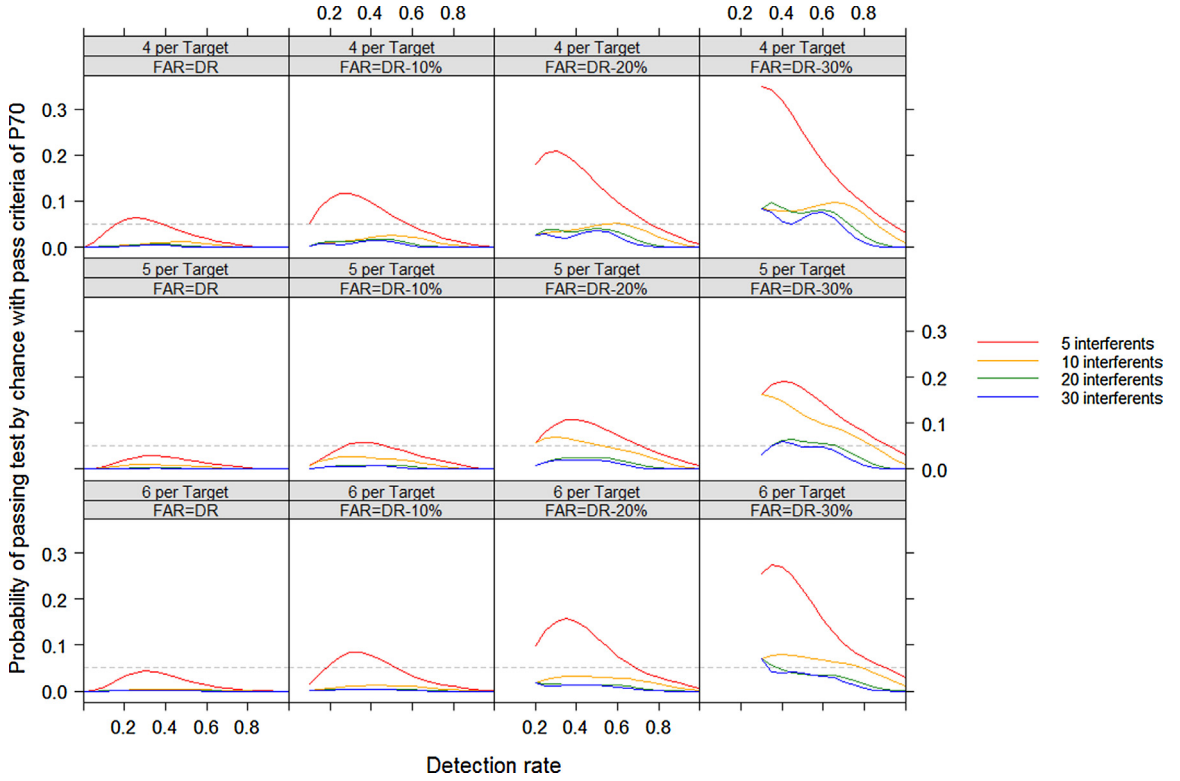

Fig. 2. Probability of passing test by chance with pass criterion of P70 (detection rate at least 70% greater than false alarm rate (FAR)) for 10–25 interferent samples per target for dogs with 0% (chance), 10%, 20% and 30% true detection rate. Graphs show that for four to six repeats per target, 10 interferents give alpha $P < 0.05$ for dogs with 0–10% true detection rate, while 20 interferents are required to give alpha $P < 0.05$ for dogs with up to 20% true detection rate, and neither 20 nor 30 interferents are able to give alpha $P < 0.05$ for dogs with 30% true detection rate.

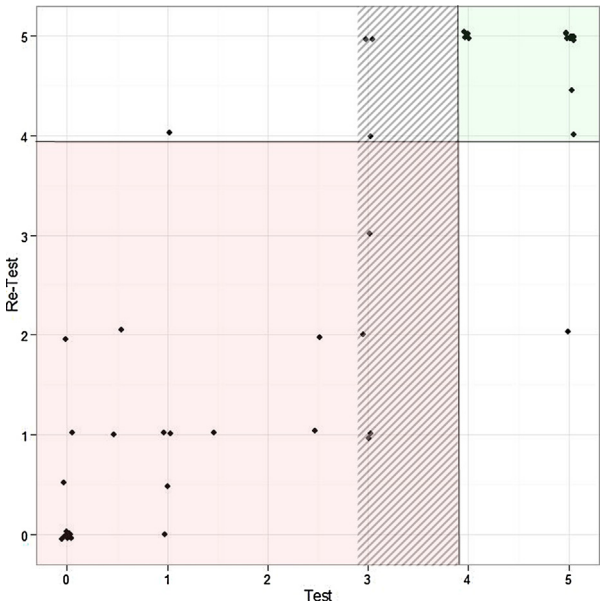

Fig. 3. Test re-test outcomes on internal validation target separated into pass (green) and fail (pink) categories. Graph shows good repeatability of poor scores (<3) and good scores (≥4). Hatched areas outline dogs with a medium score (3–3.5) on the first test and highlights potential improvement in repeatability if these dogs were required to repeat the test for accreditation purposes ($n = 51$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

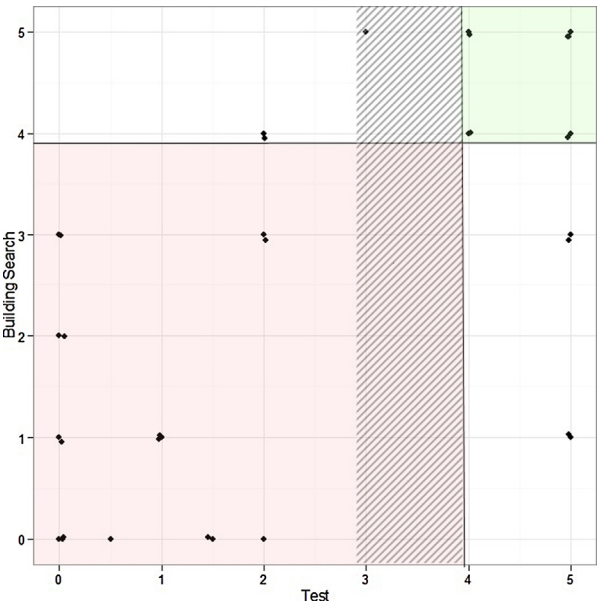

Fig. 4. Odour ID vs. building search outcome on external validation targets separated into pass (green) and fail (pink) categories. Graph shows good repeatability of poor scores (<3) and good scores (≥4). Hatched areas outline one dog with a medium score (3–3.5) ($n = 46$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the number of samples searched by all dogs following a truncated search (see Section 2.7.1).

The programme has a user interface allowing selection of dog quantity (1–3 to allow all combinations), number of targets per session (1–8), number of repeats per target (1–12 to give additional utility) and the desired test pass mark (1–12). The spreadsheet is free to use, please contact the corresponding author.

3.5. Validation

3.5.1. Internal validation: repeatability

Thirteen EDDs and six DDDs completed a total of fifty two test re-test events across five target types referred to as "internal validation targets" ($n = 13 \times$ target 1, $6 \times$ target 2, $14 \times$ target 4, $13 \times$ target 5 and $6 \times$ target 7); 97.5% of these events resulted in dogs being classified the same on both occasions (i.e. pass twice or fail twice). There was good overall repeatability using combined data from all targets (kappa, 0.80, Fig. 3).

On twenty three occasions dogs had a poor outcome on the test (<3), 22 of these dogs then failed the re-test; we can therefore be 95% confident that at least 78% of dogs with a poor odour ID test result will fail a re-test (exact 95% confidence, 0–22% pass). On twenty two occasions dogs had a good outcome on the test (4 or 5) and 21 of these dogs then passed the re-test; we can therefore be 95% confident that at least 77% of dogs which pass an odour ID test will pass a second test (exact 95% confidence, 77–100%). On seven occasions dogs had a medium outcome, scoring 3 or 3.5 (classed as a fail), four of these dogs then failed the re-test and three passed the re-test; we therefore have no confidence in the ability of the test to predict the future outcome of these dogs (exact 95% confidence, 10–82% pass). It is likely that requiring these dogs to repeat the test rather than automatically failing them would improve the repeatability of the test.

3.5.2. External validity: comparison of odour ID and search performance

Twenty six EDDs completed a total of forty six test re-test events on four targets referred to as "external validation targets" ($17 \times$ target 3, $17 \times$ target 6, $6 \times$ target 1, $6 \times$ target 5); 87% of these resulted in dogs being classified the same on both occasions (Fig. 4).

There was good overall external repeatability (kappa, 0.87). Dogs had a poor outcome on the odour ID test on 17 occasions (<3); 16 of these then failed the building search; we can therefore be 95% confident that at least 71% of dogs with a poor odour ID test result will fail a building search (exact 95% confidence, 0–29% pass). Dogs passed the odour ID test on 16 occasions (score 4–5); 14 of these went on to pass the building search; we can therefore be 95% confident that at least 62% of dogs that pass an odour ID test will pass a building search assuming that their search skills are of an appropriate level (exact 95% confidence, 62–98% pass). As search skills are a core component of conducting a search, this odour ID test must not be used in isolation, however these results indicate that odour discrimination ability shown in the odour ID test is related to odour discrimination ability shown in a building search for good and poor outcomes. Only one dog scored 3–3.5 on the test

session, confidence was therefore not calculated for this score.

3.6. Test duration

The average time taken to search six runs containing five target repeats and all associated interferent samples was 5.6 min (range 4.3–7.6 min), excluding the one off initial test set up time. Use of the software allows running orders to be generated in less than a minute.

4. Discussion

An odour discrimination test was designed that can be conducted double blind by one assessor in less than 6 min per target type. The test uses minimal materials, requires no scientific expertise to set up and is supported by a free-to-use Microsoft Excel™ programme which automatically generates test schedules. The test was found to have good internal and external validity (kappa $\geq$ 0.80) and is suggested for use by practitioners as a component of initial accreditation or ongoing quality assurance of working detection dogs.

A key requirement of any test is that it reliably measures the same behavioural component when conducted on different occasions. To the authors' knowledge, there are no published studies that explicitly measure the test re-test reliability of a canine discrimination test, although Schoon (1996) investigated the effect of olfactory discrimination test design on outcomes for a match-to-sample task. One test design required dogs to repeat the task twice and the author reported 57% correct response on the first trial with an overall correct response of 47% suggesting some loss of performance between the two trials. By contrast reliability analysis is common in the field of temperament testing, and Diederich and Giffroy (2006) note that learning by dogs during behavioural tests can often lead to poor reliability as was seen in Gazit and Terkel's (2003) study, where detection performance improved in successive search tests.

The overall test re-test repeatability of our test was high (kappa > 0.80), as was repeatability for the group of dogs that passed the first test (score 4–5) and the group of dogs that failed the first test with a low score of 0–2.5. In both cases we are 95% confident that at least 77% of the dogs would be classified the same (pass or fail) if tested on a different day. These results imply that the test is measuring a true component of discrimination ability and that it is sensitive enough to separate at least two groups of performance ability relatively accurately.

Less clear is the predicted outcome of dogs that scored 3–3.5 and were classified a medium fail. In this study seven dogs were classed as medium during the first test and three of these went on to pass the second test (i.e. changed category). Unfortunately this sample size is too small to draw any firm conclusions, however the results do suggest that there is low confidence in the reliability of pass/fail categorisation based on a score of 3 or 3.5. This suggests that the test may not be sensitive enough to accurately discriminate between three groups of ability, and may incorrectly categorise a significant minority of dogs with intermediate ability or dogs that have not had sufficient training on the process prior to testing.

This limited sensitivity should be expected given the short length of the test as there is by necessity a trade-off between test precision and test length. The shortest test length calculated to give >90% confidence in correctly assigning two groups of dogs ("poor" 40% and "very good" 95% true detection rate) as pass or fail was five repeats per target, this was therefore chosen as the length for this test. This level of precision is less than ideal, however the additional repeats required to give 90% confidence in a dog with an intermediate true detection ability, such as 75%, would make the test prohibitively long for practitioners to conduct when they have large numbers of odours to test and limited time available; increasing the test precision may therefore have an adverse effect by reducing practitioner uptake.

Possibly of even greater importance than test re-test repeatability is confirmation of the external validity of a test; this tests whether performance on the test predicts performance in the required role. There are a number of studies that test the external validity of behavioural traits as predictors of success as a working dog, however these tests are aimed at identifying potential of dogs rather than trained ability, they also require a significant amount of testing time (e.g. Svartberg, 2002; Maejima et al., 2007; Sinn et al., 2010). Rooney et al. (2007) validated a standardised search test to test trained dog ability and confirmed that overall ability recorded objectively from the test was correlated with subjective handler ratings of the recorded searches (ρ = 0.66–0.82), however as with the other behavioural tests, this approach is very time consuming and does not give information about a dog's ability to detect specific targets. In the case of contraband detection dogs the desired behaviour is the detection of a concealed target during a search of an area such as a building. This study found good overall repeatability between an odour ID test outcome and a standardised building search (kappa, 0.87). Assuming that the dogs were proficient at all other skills required to conduct a thorough building search (as determined by passing a subjective search ability test), we have confidence that 71–100% of dogs scoring <3 in an odour ID test would fail a building search for the target in question and 62–100% of dogs that pass an odour ID test would pass a building search for the target tested. Given the additional sources of variation present in a building search, the observed external repeatability was relatively high suggesting that the odour discrimination component of ability measured during the odour ID test is a significant contributor to the overall ability of a search dog.

There is a cost associated both with "incorrectly" withdrawing a good dog from service and in not withdrawing a dog with inadequate detection ability. To mitigate against both of these risks we suggest that there should be two different outcomes for dogs that fail the odour ID test. For dogs scoring 0–2.5 there is a very high probability that they would fail the test again if conducted on a different day, and would also fail a comparable building search on the odour in question; we therefore suggest that the best response would be to temporarily withdraw these dogs from service until they have successfully passed the test on the odours

in question. For dogs scoring 3 or 3.5, while they have failed to meet the pass criterion, there is limited confidence in the predictive power of the test in this range which means that they may have passed if tested on a different day. We would therefore suggest that an appropriate response would be to either repeat the test and use the total score (i.e. out of 10) to calculate the outcome with more sensitivity, or to nominally fail the dog but to give some form of concession, such as re-testing immediately or allowing a short period of time for retesting whilst continuing to work the dog. This two pronged approach mitigates against incorrectly failing dogs but also maintains the required pass standard.

Due to the significant number of additional skills required to carry out a successful building search and the requirement to find hides in various concealments, success on an odour ID test alone cannot be used to predict the success of a dog in a search scenario. Practically however success on a target in the odour ID test could be used as a basis for removing any "easy" hides containing this target from the search test or to reduce the number of hides in the search test. This would free up time to concentrate on more difficult concealments during subjective testing and reduce the discrepancy between operations where frequency of target encounter is often low, and testing where target frequency is generally very high, resulting in a more robust overall test. Use of the odour ID test also allows quantitative data on performance against each target type which is time consuming to gather in a standard search test due to the distance needed between target placements.

The aim of this study was to create a test with the lowest possible logistic burden. No dogs failed on excessive false alarms during validation tests, and false alarms are therefore excluded from analysis, however the short length of the test and the ability to truncate searches following a correct indication are possible due to pass criteria based on elements of signal detection theory. In signal detection theory each dog is scored based on a combination of their detection and false alarm rate (in this case, detection rate must be at least 70% greater than false alarm (FA) rate). In this framework, the probability of passing by chance is determined solely by samples searched and the pass criteria; no assumption is made concerning a dog's decision making process such as whether it will choose to indicate on every run regardless of target detection. This means that the number of runs and the length of each run are irrelevant which in turn allows for truncated runs of different lengths resulting in fewer samples searched. The quickest way to conduct a test would therefore have been to search multiple targets over one or two runs, however pilot testing showed that while this approach did not affect detection rates, it was logistically difficult to conduct and resulted in errors; a maximum of one target per run was therefore included in the test.

In order to maintain standardisation across dogs conducting truncated searches a Microsoft Excel™ programme was written which automatically generates standardised, balanced running orders so that all dogs search an average of 22.5 interferent odours per target (±2.5). This has the advantage of standardising test set up and notation and allowing practitioners to rapidly set up and conduct a test with no scientific input following a brief initial training session.

As access to contraband samples is restricted, the final component in reducing logistic burden is to minimise the number of samples of each target required for testing. While the test can be smoothly conducted with two samples, this is not necessarily testing ability on the target per se, rather just the two test samples. During both internal and external validation the reported repeatability was based on conducting an odour ID test with two samples and the second test with five different samples. The high repeatability of these experiments implies that performance on two targets is significantly correlated with performance on a larger number of targets. As dogs must be tested on samples that they have not previously encountered, limiting the number of samples required for testing to two will make this requirement significantly easier to achieve.

5. Conclusion

A short odour discrimination test was created that has utility for working dog practitioners either as a component of annual or biannual accreditation, or as a standalone quality assurance check in between these accreditation points. The odour ID test is sufficiently sensitive to accurately separate two groups of dogs by ability (good and poor) and mitigation approaches are suggested to reduce the probability of incorrectly assigning dogs of intermediate ability to the wrong pass/fail category. The test has high internal and external repeatability and a low logistic burden.

Conflicts of interest

The authors report no conflicts of interest.

Acknowledgments

This research was carried out by the Defence Science and Technology Laboratory [dstl] under collaborative funding provided by a number of government departments including the Home Office, Department for Transport, the Ministry of Defence and the Centre for the Protection of National Infrastructure. The funding bodies requested publication of the article but were not involved in experimental design, data analysis, interpretation or article preparation. Significant support was provided by Stephen Craigs (Durham Police), John Codd (Dyfed Powys Police) and Robert Gray (UK Border Force) and by UK police forces via access to operational handlers and dogs; support was also provided by UK Border Force via access to operational dogs and staff.

References

Bird, R.C., 1997. Examination of the training and reliability of the narcotics detection dog. Ky. Law J. 85, 405.

Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413.

Diederich, C., Giffroy, J., 2006. Behavioural testing in dogs: a review of methodology in search for standardisation. Appl. Anim. Behav. Sci. 97, 51–72.

Ehmann, R., Boedeker, E., Friedrich, U., Sagert, J., Friedel, G., Walles, T., 2012. Canine scent detection in the diagnosis of lung cancer: revisiting a puzzling phenomenon. Eur. Respir. J. 39, 669–676.

Gazit, I., Terkel, J., 2003. Explosives detection by sniffer dogs following strenuous physical activity. Appl. Anim. Behav. Sci. 81, 149–161.

Johnen, D., Heuwieser, W., Fischer-Tenhagen, C., 2013. Canine scent detection – fact or fiction? Appl. Anim. Behav. Sci. 148, 201–208.

Lin, H.M., Chi, W.L., Lin, C.C., Tseng, Y.C., Chen, W.T., Kung, Y.L., Lien, Y.Y., Chen, Y.Y., 2011. Fire ant-detecting canines: a complementary method in detecting red imported fire ants. J. Econ. Entomol. 104, 225–231.

Maejima, M., Inoue-Murayama, M., Tonosaki, K., Matsuura, N., Kato, S., Saito, Y., Weiss, A., Murayama, Y., Ito, S., 2007. Traits and genotypes may predict the successful training of drug detection dogs. Appl. Anim. Behav. Sci. 107 (3-4), 287–298.

Martin, P., Bateson, P., 1986. Measuring Behaviour. Cambridge University Press, Cambridge UK.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 http://www.R-project.org/

Richards, K.M., Cotton, S.J., Sandeman, R.M., 2008. The use of detector dogs in the diagnosis of nematode infections in sheep faeces? J. Vet. Behav. Clin. Appl. Res. 3, 25–31.

Rooney, N., Gaines, S.A., Bradshaw, J.W.S., Penman, S., 2007. Validation of a method for assessing the ability of trainee specialist search dogs. Appl. Anim. Behav. Sci. 103, 90–104.

Schoon, A., 1996. Scent identification line-up by dogs (*Canis familiaris*): experimental design and forensic application. Appl. Anim. Behav. Sci. 49, 257–267.

Sinn, D.L., Gosling, S.D., Hilliard, S., 2010. Personality and performance in military working dogs: reliability and predictive value of behavioural tests. Appl. Anim. Behav. Sci. 127 (1–2), 51–65.

Svartberg, L., 2002. Shyness-boldness predicts performance in working dogs. Appl. Anim. Behav. Sci. 79 (2), 157–174.

Williams, M., Johnston, J.M., 2002. Training and maintain the performance of dogs (*Canis familiaris*) on an increasing number of odor discriminations in a controlled setting. Appl. Anim. Behav. Sci. 78 (1), 55–65.