SCENT IDENTIFICATIONS BY DOGS (CANIS FAMILIARIS): A NEW EXPERIMENTAL DESIGN

by

G.A.A. SCHOON[1)]

(University of Leiden, Department of Criminalistics and Forensic Science; Institute of Evolutionary & Ecological Sciences, ethology group; National Dutch Police Tracker Dog Centre; PO Box 9520, 2300 RA Leiden, The Netherlands)

(Acc. 19-XI-1996)

Summary

Traditionally, match-to-sample designs are used for forensic scent identifications: the scent of a perpetrator on a *corpus delicti* is matched to the scent of a suspect. In a number of cases, *e.g.* when the suspect is innocent, no match is possible, which leads to specific difficulties. In a new design an odd-even paradigm was followed, and care was taken to meet forensic prerequisites in the experimental setup. Four dogs were trained to compare a human scent (odour 1) on stainless steel tubes, training objects, or typical forensic objects to a human scent on a stainless steel tube (odour 2). Comparisons could be either 'odd' ($1 \neq 2$) or 'even' ($1 = 2$). If the dogs performed poorly in the beginning of an experimental series, they were disqualified from making forensically interesting comparisons. Realistic experiments demonstrated the ability of dogs to compare scents following this protocol, but also showed that the results were influenced by the type of odour 1 and by the type of comparison. The performance of the dogs is compared to the performance of operational dogs in a match-to-sample design: the level of matching 'even' scents is comparable, but the level of non-matching in 'odd' comparisons is substantially higher in the new design. Scent identifications following an odd-even paradigm seem to be more reliable than the customary design. Introducing this new design would however require significant changes in attitude and working conditions of the police.

[1)] This project was made possible by the Dutch Police Force who assigned two dog handlers to work with me and train the dogs. I would like to thank these two handlers, Hans Evers and René Timmerman, for their creative input and dedication. I would also like to thank prof. dr. J.M.H. Vossen, prof. dr. P. Sevenster and prof. dr. C.J. ten Cate for their contributions throughout the project and their appraisal of the manuscript.

Part 1. Description of design and legal justification

1.1. Introduction

From the beginning of this century police have used the remarkable olfactory acuity of dogs for their investigations. The most controversial of these abilities has been the ability of dogs to match the scent of a perpetrator on an object related to a crime (the *corpus delicti*) to the scent of a suspect. Statistic testing of this ability shows that dogs are capable of matching scents under very different circumstances (Kalmus, 1955; Hepper, 1988; Sommerville *et al.*, 1990; Brisbin & Austad, 1991; Toner & Miller, 1993; Settle *et al.*, 1994; Schoon & de Bruin, 1994; Schoon, 1996). However, when a single 'scent identification', indicating that the suspect is indeed the perpetrator, is presented as a piece of evidence in court the reliability of each single match becomes a matter of importance.

In the course of an investigation into the reliability of scent identifications in the Netherlands a number of problematical issues were identified that influenced the reliability of these identifications. In summary, the most important issues were:

* the results varied from identification to identification. Possible causes are varying motivation of the dog or varying physiological ability (olfactory acuity varies from day to day and can be affected by illness, hormonal state, drugs, *etc.*);
* the way of working with the dogs in the line-up was not very systematic;
* the customary experimental protocol did not exclude handler-influence on the dog;
* the scent of some people seemed attractive for some dogs;
* the scent of some people led to systematically poorer results than average;
* when no matching scent was available in the identification-line-up the dogs had to inhibit all responses (go-no go paradigm) whereas they were trained primarily to respond to one of six scents in the lineup (match-to-sample paradigm). This could cause confusion;
* according to the customary experimental protocol (for a description, see Schoon, 1996), the dogs respond by retrieving a hand-scented stainless steel tube, and are rewarded with a retrieving game using this

same tube. The distinction between response and reward is therefore minimal, especially since the tubes are not fixed in any way and can be picked up freely.

Although some of these issues could be addressed by slight variations in the experimental protocol and identification method (Schoon, 1996; in prep.), it was felt that a new approach could solve more of these issues.

1.2. Prerequisites of a new approach

Any new approach should meet legal prerequisites and at the same time make the experimental protocol more straightforward. It should also be more reliable than the customary method, taking into account the customary judicial opinion that a false accusation of an innocent person is worse than releasing a guilty one.

Taking the problems described in the introduction into account, a new approach should meet the following points:

1) the dog should be willing and capable of working at the time of the scent identification, preferably one should know how willing and capable the dog is at that moment;
2) a) the dog should not have any prior preference for the scent of the suspect,
   b) but should not have problems with identifying this specific person on his/her scent either;
3) there should be a standardised way of working;
4) there must be no influence of the handler on his dog when the dog is choosing;
5) the dog should be able to respond in some way when there is no matching scent, or when it is uncertain whether the odours match;
6) the method should offer more distinction between response and reward.

1.3. Experimental design

An experimental protocol based on an 'odd-even' paradigm was designed to meet these points. In each experiment, the dog was presented with a series of trials. Each trial consisted of a comparison of two odours, and the dog learned to respond in one way if the odours came from the same

person, and respond in another if they did not. In the series, all choices but one (the forensic interesting comparison of the odour on the *corpus delicti* and the odour of the suspect) were comparisons with a 'known' outcome, and the performance of the dog in these trials could be seen as indicative of his reliability on the comparison of the odour on the *corpus delicti* and the odour of the suspect.

As response, retrieving a stainless steel tube was chosen. Two tubes were fixed on a 50×120 cm platform with 40 cm between them. The first step in the protocol was that the dogs were given the scent of a person on an object by their handler (odour 1). They were trained to then go to one tube on the platform (for each dog: always the same side first). This tube was also scented (odour 2). If the odours matched (an 'even' trial where odour 1 = odour 2), the dogs were trained to respond to this tube. If the odours did not match (an 'odd' trial where odour 1 $\neq$ odour 2), the dogs were trained to go to the other tube on the platform, a 'blank' unscented tube, and respond to this blank tube. When the dogs responded, the handler (who was unaware of the correct choice) would signal by raising his hand. If the dog was responding to the correct tube, the experimenter (who could only observe the dog on a video screen) would release this tube for the dog to retrieve. If the dogs choice was incorrect, the tube was not released and the dog was recalled. A correct response was also followed by some petting, while after a false response the dogs were not petted. A schema of the setup is given in Fig. 1.

The experimental protocol followed in this study consisted of 12 trials, approximately half of these were 'even', the others 'odd'. In the first 6 trials, points 1 and 2a were addressed: the performance level of the dog was established, and by using the scent of the real suspect in an 'odd' trial the lack of special interest of the dog in this persons scent was demonstrated. If the dog made a single error in an 'odd' trial (thereby incorrectly responding to a scented tube) the dog was disqualified. If the dog made two errors in an 'even' trial (thereby incorrectly ignoring the scented tube) the dog was also disqualified. This difference in evaluating the mistakes was inspired by the thought that the gravity of the mistakes differed. Incorrect responses to a scented tube can lead to an incorrect accusation of the suspect in the 'forensic' trial and must be prevented at any cost. The dog must be absolutely sure that odour 2 originates from the same person as odour 1. If

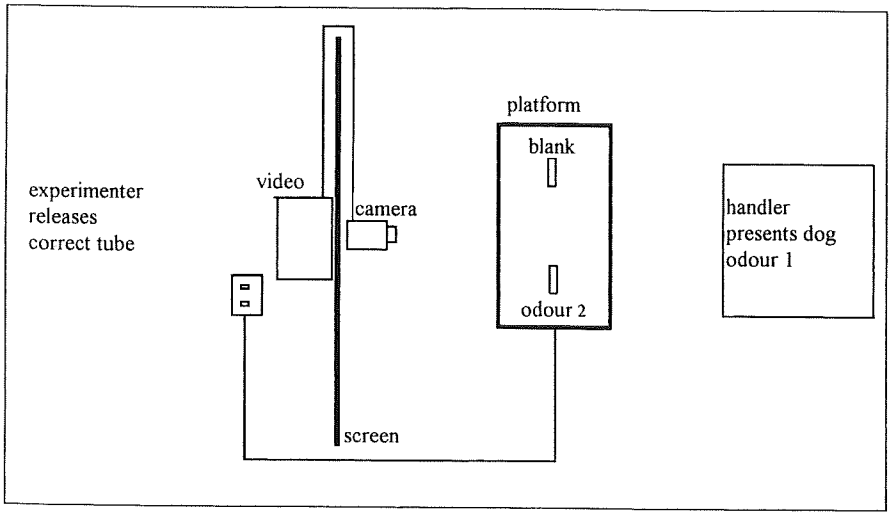

Fig. 1. Illustration of experimental set-up.

the dog is not sure, or when it considers the odours to be clearly different, it should select the blank tube. Incorrectly ignoring a scented tube can be the result of a number of (physiological or motivational) reasons we do not have control over, and only leads to missing a culprit which is a less serious mistake than a false accusation.

In the next 5 trials, the *corpus delicti* was presented twice as odour 1: once in an 'odd' trial, once in the 'forensic' trial with the odour of the suspect. In this way the handler could not be certain which trial was the actual 'forensic' trial, and this prevented possible influence of the handler on his dog in the direction of false accusations (an aspect of point 4). In the last choice, the dog was checked for his ability to work with the scent of the suspect in an 'even' trial with the scent of the suspect on an object as odour 1 and on a tube as odour 2 (point 2b). An example of a series is given in Fig. 2.

In summary: a complete scent identification test consisted of 9 choices with different objects selected from the array of training objects or tubes as odour 1 ('standard training choices'), 2 choices with a typical *corpus delicti* as odour 1, and a last choice also with a training object as odour 1. Ten choices in the test had 'known' answers that could either be 'even' or 'odd'. Only the correct responses were rewarded. One choice was the

Trial	Odour 1	Odour 2	Correct/ Wrong
1	**A**	**G**	
2	**B**	**B**	
3	**C**	**C**	
4	**D**	**D**	
5	**E**	**suspect**	
6	**F**	**F**	
7	**corp.delicti**	**H**	
8	**G**	**I**	
9	**H**	**H**	
10	**corp.delicti**	**suspect**	
11	**I**	**J**	
12	**suspect**	**suspect**	

Fig. 2. Example of an experimental series of 12 trials. A-J and the suspect are all different people.

forensic trial and was always rewarded. The last choice was always an 'even' choice.

This two-choice protocol is very simple in comparison to the customary protocol where the dogs have to choose between 6 tubes. It led to a very consistent way of working (point 3) since the dogs always began on the same side of the platform. The dogs worked unleashed, and they made a direct choice for the scented tube or for the blank one which left no room for influence by the handler during the choice (point 4). The dogs could always be rewarded in the same way if they performed correctly (point 5) by retrieving a tube. The difference between the response of the dog (scratching or biting at the chosen tube) and the reward (retrieving the tube, being petted) and the time delay between the two created a distinction between response and reward (point 6). The response of the dog was defined by the tube the dog was responding to at the moment the handler signalled. This was done to prevent ambiguity: a person knowing the correct choice could interpret responses 'in favour' of the dog. By letting the handler, who did not know which tube was the correct choice in each trial, determine the moment the dog had chosen any influence of the experimenter releasing the tubes (who of course did know the correct choice) was prevented. Any other influence was prevented by not letting the experimenter observe

the dog directly when on the platform but through a video connection, as illustrated in Fig. 1.

Part 2. Realistic experiments

To evaluate the method, four dogs were trained in the above way and subjected to a series of experiments. A typical forensic case was the model for the experiments: typically crime related objects were used as *corpus delicti*, typically found 1-2 weeks earlier than the suspect was caught, typically the suspect being the only one staying in a cell and the other scents in an experiment being scents of police officers.

2.1. Material and methods

Animals

Four dogs were trained by two handlers (each handler trained two dogs) during approximately 1 year. Three dogs (dogs 1, 2 and 3) were Malinois shepherd females, just one year old when the training started and without previous training. The fourth dog (dog 4) was a 6 year old Malinois shepherd male, he was a trained police dog. Police dogs are not extensively trained in tracking, but are trained to locate human-scented objects.

Training/experimental area

The training and the experiments were done in an unused garage block of a Dutch police station. The temperature could not be regulated: it varied from −5°C in winter to more than 30°C in summer. During very cold weather the objects and tubes were kept in front of an electric heater prior to being used for a trial. The area was approximately 15 m long and 5 m deep, and had a coated concrete floor that was regularly cleaned.

Training

In general, the steps followed traditional training methods where possible. The dogs were first trained to retrieve stainless steel tubes. In the set-up, the will to retrieve a stainless steel tube was the leading motivation for the dog, and being allowed to retrieve one was to be the reward, therefore dogs were chosen that really enjoyed this.

Teaching the dogs in the odd-even paradigm was done in three general stages (write to the author for more details):

Stage 1. Teach the dog to search for the matching scent. — The dog is systematically given a human-scented object to smell at (as odour 1 in Fig. 1). It is taught to search for the matching human scent on a tube and to ignore non-matching scents;

Stage 2. Teach the dog to retrieve a blank tube if there is no matching scent. — The dogs have learned to work systematically: they spontaneously develop a strategy where they only smell at the first tube and if this is a non-matching odour they 'grab' the second tube without smelling it, so this was relatively easy to train;

Stage 3. Consolidate the training and expand the odour experience of the dog. — Typically crime-related objects like dirty screwdrivers, guns, knives, scent samples from car seats, *etc.* were introduced in the training, and the length of time these had been scented was varied. The time between scenting the objects and the matching tubes was also increased: in training situations they were scented simultaneously but in actual crimes objects are scented earlier, when the crime is being committed, than the tubes, which are scented once the police have a suspect.

From time to time the dogs seemed to be following alternative strategies for their choices. One dog favoured alternation, once a 'win stay, lose shift' strategy was detected, another dog seemed to group all new objects into a separate category and always chose the scented tube in those cases, another dog went through a phase in which he always chose the blank tube if the object presented as odour 1 was a tube, *etc.* When such a strategy was detected, it was counterbalanced in the training by provoking the mistakes and reprimanding the dog when he made a mistake.

Besides training the dogs on the two-choice platform, in this stage the dogs were also trained on a six-choice platform where they were always given 'even' trials. The purpose of this training was twofold: to 'break' alternative strategies, and to accelerate the expansion of experience of the dogs. This training gave the dogs a more difficult task (instead of comparing two odours, now compare 7), requiring more patience (6 tubes to smell instead of 1) and within a trial a longer period of concentration, while the total training time was a lot shorter (10 instead of 45 minutes). Alternative strategies did not develop here, the dogs really had to use their noses. This 6-choice training was done approximately once a week.

There were three guidelines throughout the training. The first was to *build up* the training in a classical 'shaping'. The dogs first had to master one step before the training went on to the next step, and within a step the dog first had to do something 'easy', something he already could do well (*e.g.* make a comparison based on a metal object), followed with a 'difficult' choice (*e.g.* make a comparison based on a wooden object) using the scent of the same person as in the 'easy' comparison.

The second guideline was to *step down* whenever a dog made a mistake. After a first mistake on a given level the choice could be repeated using the scent of a different person: sometimes dogs seem to have difficulty with the scent of a particular person. If the dog made a second mistake, the training continued at a lower step, and immediately started to build up again.

The third guideline was a carefully *differentiated reward/punishment scheme.* A correct response was always rewarded with retrieving the tube and petting. In stages 2 and 3, there are two kinds of mistakes the dogs could make: they could choose the scented tube in a 'odd' trial, which is a grave mistake since it could lead to a false identification in a forensic test, and they could choose the blank tube in an 'even' trial which meant missing a possible identification which is a lesser offence in our judicial system. When the dogs made a mistake in a 'odd' trial a beep was sounded, and this was followed by a vocal reprimand for the dog by his handler. The beep alone became a strong signal for the dogs to discontinue responding and return to the handler. When the dogs made a mistake in an 'even' trial they were recalled in a normal tone by the handler and they were not rewarded. This was a neutral response. Although the dogs themselves of course were unaware of the 'value' of their mistakes, we wanted to minimize mistakes in 'odd' trials and teach the dogs to only choose odour 2 if they were absolutely certain this was the same as odour 1.

Experiments

For the experiments, *corpora delicti* were prepared by two civilians working at a police station. They acted as perpetrators (A and B in Table 1), and the *corpora delicti* were prepared in a 'realistic' way: objects were carried in the pocket for approximately 15 minutes, and handled for about 5 minutes. These objects were screwdrivers, spanners, and plastic pistol buttplates. In addition, the 'perpetrators' wore sweatshirt cuffs for 15 minutes and handled them for another 5 minutes. Scent samples were collected from the seat of the car they had driven to work in, both lived approximately 30-45 minutes away from work. The scent samples were taken by putting a 10 × 15 cm scent collecting cloth (cotton bandage material) on the seat of the car, covered with aluminium foil, and left there for approximately 2 hours. The pistol buttplates, sweatshirt cuffs and scent samples were collected in glass jars with twistoff tops, the screwdrivers and spanners were collected in plastic bags, as is customary police protocol.

In the week after the preparation of the *corpora delicti* (7-12 days later), the scent identification material needed for the experiments was prepared. This is a realistic time-span: in police inquiries, two-thirds of the scent identifications are done within 2 weeks after the crime had been committed. The experiments were divided into two groups: half were experiments in which the perpetrator (A) was also the suspect, these were potentially 'even' experiments, and half were experiments where the perpetrator (B) was not the suspect (C), so these were 'odd' experiments.

The perpetrator/suspect (A) and the innocent suspect (C) were both civilians working at a police station. The other 10 odours were used only once per dog, and came from students of a police training school in another town. This simulated police reality as well: usually the suspect is the only one in a cell and the other odours belong to police officers. All persons scented training objects by handling them during 1 minute and in their pockets for another 5 minutes, and tubes by handling them 5 minutes. Objects and tubes were collected separately in two glass jars with twist-off tops.

Each dog was given 10 experiments: 5 'even' and 5 'odd' with each of the 5 different kinds of *corpora delicti*. The handlers knew the *corpora delicti* were 2 weeks old and knew the experiments could be either 'even' or 'odd' but they did not know the ratio between these two. The dogs were given one or two experiments a week on previously designated days. On other days the training continued. In the experiments, odour 1 that was presented to the dog was a scented training object approximately half of the time, and a scented stainless steel tube the other half.

2.2. *Results*

In Table 1, the results of the four dogs are given. Two females were disqualified twice, one thrice, and the male was disqualified four times. On average, the suspect was correctly identified in 4 out of 14 experiments (28.6%), and incorrectly identified in 1 out of 15 experiments (6.7%).

Analysing the results of the 'known' choices where the dogs were qualified, some differences between the dogs become clear. Dog 1 performed all standard training choices correctly in 5 of the 10 experiments, dog 2

TABLE 1. *Results of experiments with realistic corpora delicti in realistic situations*

	Suspect A = *Perpetrator* A					*Suspect* C ≠ *Perpetrator* B				
	buttplate	screwdriver	spanner	cuff	scent sample	buttplate	screwdriver	spanner	cuff	scent sample
Dog 1	$X_{10,0,0}$	$0_{9,1,0}$	$0_{10,0,0}$	D	$0_{10,0,0}$	$0_{10,0,0}$	$0_{10,0,0}$	$0_{9,1,0}$	$0_{9,0,1}$	D
Dog 2	$0_{8,2,0}$	D	$0_{10,0,0}$	$0_{8,2,0}$	$X_{8,1,1}$	$0_{9,0,1-}$	$0_{10,0,0}$	$0_{10,0,0}$	$0_{8,1,1}$	D
Dog 3	D	D	$X_{10,0,0}$	$X_{9,0,1}$	$0_{9,1,0}$	$0_{10,0,0}$	$0_{10,0,0}$	$0_{10,0,0}$	$0_{9,0,1}$	D
Dog 4	$0_{9,0,1}$	D	$0_{9,1,0-}$	$0_{8,2,0}$	D	$0_{10,0,0}$	$0^{*}_{9,1,0}$	D	$X_{9,1,0}$	D

* Other suspect, same perpetrator.

D: disqualified.

X: chooses suspect in forensic test (correct choice in suspect A = perpetrator A).

0: chooses blank in forensic test (correct choice in suspect C ≠ perpetrator B).

Subscript: results in 10 'known' choices: N correct, N false rejections, N false identifications, followed by − if the last choice was incorrect.

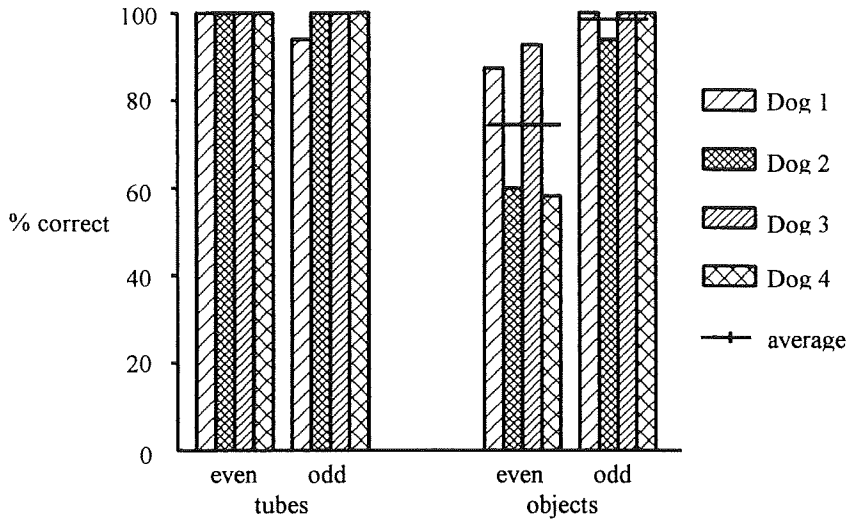

Fig. 3. Results of the standard training choices per dog, divided into odour 1 'tubes' or standard training 'objects', and into 'even' and 'odd' choices of each.

in 3 out of 10, dog 3 in 4 out of 10 and dog 4 only once. In Fig. 3 the choices with the training objects and the tubes as odour 1 (the 'standard training choices') are divided into 'even' and 'odd' choices for each of the dogs.

Choices based on a tube as odour 1 rarely led to incorrect responses: only 1 out of 145 choices was an incorrect identification. Choices based on training objects as odour 1 led to on average 1.7% incorrect identifications in 'odd' trials, with very little difference between the dogs. Choices based on training objects led to on average 24.6% incorrect rejections in 'even' choices, but dogs 1 and 3 made far less incorrect rejections (12.5 and 7.1% respectively) than dog 2 and dog 4 (40.0 and 41.7% respectively). However, these differences were not significant (χ^2, df = 3, $p = 0.061$). Dogs 2 and 4 were the only ones who also each once missed the final identification of the suspect (the '–' in Table 1).

Since the results of the dogs did not differ, the results could be combined for further analysis. Combining the results of the odour 1 'tubes', 'training objects' and 'forensic objects' for all dogs it is clear that for the 'even' choices, the type of odour 1 has a significant effect on the results, with 'tubes' leading to the best and 'forensic objects' to the poorest results. For the 'odd' choices it seems that more mistakes are made with the 'forensic

TABLE 2. *Summarised results of the experiments where the dogs were qualified, divided into different odour 1 objects*

	Tubes	Training objects	Forensic objects	Significance
Even choices				
Correct	81	43	4	
Incorrect	0	14	10	$p < 0.0001$
Odd choices				
Correct	63	58	52	E too small,
Incorrect	1	1	6	not possible

Differences were tested using χ^2, df = 2.

objects', but this effect cannot be tested due to insufficient data (Table 2). The pooled results also show that the dogs made significantly more mistakes in the 'even' comparisons than in the 'odd' comparisons for the training and forensic objects (χ^2, df = 2, $p < 0.0001$).

Analysing the incorrect choices in Table 2, a division was made into choices where the dog chose the 'same side' when he should have chosen the other side as the previous choice, and choices where the dog 'changed sides' when he should have chosen the same side as before. This showed that 20 of the 32 mistakes were 'change side' situations ($p < 0.05$, T-test). Only one dog (dog 2) had regularly alternated during the training, but dog 4 also demonstrated a willingness to change sides.

2.3. Discussion

The results of the realistic experiments lead to the following conclusions:

a) dogs are capable of working with scents in an odd-even paradigm;
b) the type of object presented as odour 1 significantly influences the result in 'even' comparisons;
c) the type of object presented as odour 1 does not seem to influence the result in 'odd' comparisons.
d) the dogs make more mistakes in 'even' comparisons than in 'odd' comparisons.

Ad. a)
Although the dogs are capable of working in an odd-even paradigm, they did develop alternative strategies during the training, as was described earlier. Training the dogs for a longer period, and incorporating more

training on the 6-choice platform where alternative strategies are not easily developed, might lead to better results. This idea is supported by the fact that the dog making the least mistakes, dog 1, was in training for 2-3 months longer than the other dogs, and by the experience that alternative strategies were at least temporarily broken by training on the 6-choice platform.

Ad. b)

The differences between the dogs were not significant, but only just so. Looking at the results of the dogs 1 and 3, they showed no difference in performance level between tubes or standard training objects as odour 1, whereas dogs 2 and 4 showed a big difference in performance. The behaviour of these two dogs illustrates some further problems.

Dog 4 had been in training with his handler as a policedog for 4 years. The handler described his dog as a very eager worker who reacted strongly to him. His experience was that the dog did not like to make mistakes. The behaviour of the dog after an incorrect choice in the odd-even paradigm could be interpreted as confirming this: when the dog made a mistake in a trial, he would not go to the platform for the next choice as readily as usual, but needed to be stimulated to do so. Perhaps the difference in punishment of the mistakes (a 'non-recognition' mistake being not rewarded, but a 'false accusation' mistake being scolded) prompted the dog to prefer the 'non-recognition' mistake and therefore only choose the scented tube when he was very sure. This implies that the dog found it more difficult to compare scented objects to scented tubes than to compare scented tubes to scented tubes, which led to the lower level of correct identifications with the training objects.

Dog 2 needed extra stimulation to continue working: she often started well and made the first choices readily, but after 5-6 choices she seemed no longer motivated and the readiness to go to the platform decreased, and her concentration seemed less since she made more mistakes. This can also be seen in the results of the experiments: although she was only disqualified twice by mistakes in the first six choices of each experiment, she had on average the highest number of mistakes per experiment. This dog also was seen to alternate quite regularly and it was sometimes difficult to see if she was really working well or whether good results were an accidental consequence of her alternating scheme fitting in well with the previously determined sequence of choices. However, the good results with

the tube-tube comparisons show that any kind of other strategy, or lack of motivation/concentration, only emerged with object-tube comparison, supporting the conclusion that these comparisons are more difficult.

The results with the standard training objects of dogs 1 and 3 are quite good (87.5 and 92.9% respectively). The introduction of a selection criterion in the training seems necessary to obtain good results. Further training on forensic-style objects should show how high the recognition-performance can be. Following observations with humans (Walk & Johns, 1984), the performance level in a comparison of two odours should be higher than the performance level in a comparison of multiple odours, as in the customary match-to-sample paradigm used for scent identifications.

Ad. c)

The level of mistakes in the odd trials show a low percentage of mistakes (<2%) for the tubes and the standard training objects as odour 1. However, for the forensic objects the level of mistakes in odd choices is higher (Table 2: average 10.3%; per dog: dog 1 0/16 mistakes, dog 2 2/16 mistakes, dog 3 2/14 mistakes, dog 4 2/12 mistakes). This could well be a result of incomplete training. Training on forensic style objects only started in stage 3 of the training, and the ratio of forensic style objects *vs* standard training objects and tubes as odour 1 was low, on average 1.5 : 10. The reason for this is that more difficult objects were only practised if the dog was working well, and it took some trials to establish this. A second aspect of the incomplete training was the lack of balance between 'even' and 'odd' trials with these objects. Reviewing the training of the last 20 training sessions with each dog, it appeared we had offered non-training objects 83 times in 'even' trials and 44 times in 'odd' trials. The dogs choice in these non-standard object trials was the scented tube 57 times, which was correct 52 times, and the blank tube 70 times which was only correct 39 times. This might explain a bias towards the scented tube for non-training object trials as this strategy leads to more success (91% *vs* 56%). A correct training balance should lead to the same level of mistakes for forensic style objects as for the standard objects and the tubes. Training in the 6-choice method more often could prove to be helpful since alternative strategies are less easily developed here.

TABLE 3. *Possible responses and pay-off matrix in terms of the theory of signal detection*

	Response dog	
Reality	yes = odours match	no = odours do not match
Signal present = 'even' comp.	'hit': correct identification rewarded: retrieval and petting	'miss' neutral: no retrieval and petting
Signal absent = 'odd' comp.	'false alarm': incorrect ident. punished: beep and verbal reprimand	'correct rejection': correct non-ident. rewarded: retrieval and petting

Ad. d)
This result is exactly what we tried to obtain with the differentiated reward/punishment scheme: mistakes in 'odd' trials could lead to incorrect accusations in forensic trials and had led to a beep and a verbal reprimand in training, mistakes in 'even' trials could be due to a number of causes and had led to a more neutral non-reward. This fits in well with what is known from the theory of signal detection: the differentiated reward/punishment scheme led to a pay-off matrix that enforced a cautious criterion for the dogs (see Table 3). A 'no, the odours do not match' could lead to either a neutral response or a reward, but a 'yes, the odours do match' could lead to a punishment or a reward. The net effect was to succesfully bias the dogs to a more cautious decision, as we particularly wanted to prevent false identifications.

Part 3. Evaluation of method

In evaluating the method, three aspects should be taken into account:

- does the method meet legal prerequisites?
- are the results (especially those leading to an identification) more reliable than those obtained in the customary way?
- is the method feasible on a larger scale?

The method was specifically designed to meet legal prerequisites. How the method meets each point is described extensively in paragraph 1.3. The weakest part in the protocol is the remaining contact between the handler and his dog: the handler therefore can still influence the dog. This contact is necessary since the retrieving game with the handler is the dogs reward.

The possible influence of the handler on his dog was minimal during the actual choice: the choice the dog has to make is simple and is made quickly. Observing the dogs led to the conclusion that the dogs did not pay attention to the handler while choosing. However, after a mistake or when the object presented to the dog as odour 1 was new, or perhaps carried little odour, some dogs were reluctant to leave the handler to go to the platform to make their choice. This could be the result of the net effect of the pay-off matrix (Table 3). This point merits further attention: probably each dog requires a specific pay-off matrix to keep the motivation to work high enough but the decision criterion sufficiently cautious. However, in general the method meets the legal prerequisites formulated in paragraph 1.2 well.

The approach of incorporating an experimental question with an unknown outcome in a series of experimental questions with a known outcome leads to better interpretation of their results. It is similar to callibrating a machine that measures the alcohol level in blood for example: here an occasional 'known' bloodsample is tested and the machine result is compared with the known alcohol content of the sample. Here, the experimental questions with an unknown answer are the bulk of the measurements. The experimental question with the known answer ensures that the others were done correctly. In working with animals the balance between the 'known' and 'unknown' has to favour the 'known' answers. When a trained response is used as a measurement, one has to be aware that these responses have to be maintained which can only be done in 'known' trials.

A comparison of the results obtained in this study with the customary scent identification methods can be made by recalculating results using this customary method to a choice out of two, instead of 6, as suggested by Rosenthal (1991). The formula used to do this is:

$$\pi = (p(k-1))/(p(k-2)+1),$$

where π is the result (figure between 0 and 1) in a two-choice test, p is the result (figure between 0 and 1) in the k-choice test, k is the number of alternatives in that test.

In two recently published studies with Dutch police tracker dogs (Schoon & de Bruin, 1994; Schoon, 1996) the best results were on average 58% correct in a 6-choice test, using stainless steel tubes scented in the pocket or by hand for 30 s as odour 1. These were comparable to the 'even' trials using tubes or standard training objects as odour 1 in this study as all the

material was scented on the same day. Recalculating the 58% correct leads to $\pi = 87\%$, which is comparable to the results obtained in this study by dogs 1 and 3 for the standard training objects, but lower than the average results of all dogs with the tubes as odour 1.

For the 'odd' trials there is no comparative data available, but some inferences can be made from the same studies. When the dogs did not make a correct choice in the customary 6-choice method, they would pick up nothing half the time and pick up an incorrect tube the other half. Extrapolating this 50% nothing-50% fault to factually 'odd' situations, this would mean 50% correct-50% fault. Dividing these faults over the 6 tubes in the row the chance that a particular person (*i.e.* the suspect) was incorrectly identified would be 50/6, which means approximately 8%. Recalculating this 8% according to Rosenthal's formula leads to $\pi = 30\%$. This is substantially higher than the results obtained in the 'odd' trials in this study: <2% incorrectly identified in trials with the standard training objects or tubes as odour 1, and 10% with realistic forensic objects as odour 1.

By comparison, the odd-even paradigm seems to lead to better results than the match-to-sample paradigm. The difference cannot be sufficiently reduced to a difference between the two different groups of dogs that were used: the results in the 'even' trials are very comparable, the real difference is found in the 'odd' trials. As stated in the introduction, the customary protocol mixes two training paradigms: a 'match-to-sample' paradigm that is the basis of the training, but when there is no matching scent the dogs may not respond at all, and a 'go-no go' paradigm is followed. The basic idea of the odd-even paradigm was to prevent this kind of confusion, and the better results in particularly the 'odd' choices indicate that this objective was met.

For forensic purposes the recalculations to a two-choice system are irrelevant: the chance of recognition in 'even' trials and the chance of an incorrect identification in 'odd' trials that are the direct result of the complete method are necessary to establish reliability. When using scent identifications in court, the court should know how often an identification is in reality 'true' and how often it is 'not true'. This is best described by the 'diagnostic ratio' (Malpass & Devine, 1984): the reliability of an identification is the proportion 'hits' in Table 3 divided by the proportion 'false alarms'. Based on the average results with the training objects in

this paper, the reliability of scent identifications in the odd-even paradigm is 75.4/1.7% which leads to a diagnostic ratio of 44.4, but based on the results with the realistic *corpora delicti* the diagnostic ratio is no more than 4.3. Diagnostic ratio's for eyewitness identifications vary between 9 (Cutler *et al.*, 1987) and 15 (Wagenaar & Veefkind, 1992). Recalculating data from a survey of forensic laboratory proficiency testing (Peterson & Markham, 1995) reveals that common forensic methods vary in diagnostic ratio's from 3 to 160. Courts need to be able to compare the reliability of methods in order to evaluate the evidence presented to them.

Feasibility of this odd-even paradigm has to be seen in the light of the way scent identifications are currently carried out. The Dutch situation in 1995 will illustrate this. In that year there were 14 trained dogs distributed over the country available for scent identifications. Only one site (where 2 dogs worked) had an indoor training facility, the others worked outdoors. All dogs were also trained in tracking, and in searching for humans or human-scented objects. The handlers usually also had a trained narcotics dog. Both dogs needed training, and both dogs were called upon for police work. The handlers often trained alone. Training a dog in an odd-even paradigm is not possible under such circumstances.

Training dogs in an odd-even paradigm for forensic purposes can be done on the following conditions. Indoor facilities are necessary. A plain, empty room, large enough for the dogs to run around in and to work with a videorecorder, lacking 'interesting' objects, where the temperature can be regulated and the floor can be easily cleaned is a minimum. The handlers will have to work together in small groups. These groups need to consist of people willing to work with each other in a team-spirit, trusting each other and valuing each others opinion. Regular training is essential and should have the highest priority, therefore a combination with irregular police work (like a narcotics dog) seems impossible. Collecting odours and the cleaning of material is a substantial task that requires facilities and planning. The handlers need to be able, and willing, to train dogs in a reward-oriented way. They need to be open-minded, and willing to learn about odours, olfaction, and learning-theory. An animal psychologist or ethologist versed in learning behaviour should be responsible for the link to scientific developments, provide for the education of the handlers and supervise the learning process of the dogs, as they tend to develop alternate

strategies. Selection of the dogs should be more stringent, and should focus on the level of reliability as well as on stamina for the long training and experimental sessions.

In summary, scent identifications following an odd-even paradigm meet forensic prerequisites and seem to be more reliable than scent identifications following a customary protocol. Proper training of the dogs requires conditions that are currently not met, and will require a significant change in attitude and working-method on the part of the handlers and the police in general. Assessing cost and benefit should show how viable this new approach is.

References

Brisbin I.L., Jr., & Austad, S.N. (1991). Testing the individual odour theory of canine olfaction. — Anim. Behav. 42, p. 63-69.

Cutler, B.L., Penrod, S.D. & Martens, T.K. (1987). The reliability of eyewitness identification: the role of system and estimator variables. — Law and Human Behav. 11, p. 233-258.

Hepper, P.G. (1988). The discrimination of human odour by the dog. — Perception 17, p. 459-554.

Kalmus, H. (1955). The discrimination by the nose of the dog of individual human odours and in particular of the odours of twins. — Br. J. Anim. Behav. 5, p. 25-31.

Malpass, R.S. & Devine, P.G. (1984). Research on suggestion in lineups and photospreads. — In: Eyewitness testimony: psychological perspectives (G.L. Wells & E.F. Loftus, eds). Cambridge University Press, New York.

Peterson, J.L. & Markham, P.N. (1995). Crime laboratory proficiency testing results, 1978-1991, II: Resolving questions of common origin. — J. Forensic Sci. 40, p. 1009-1029.

Rosenthal, R. (1991). Meta-analytic procedures for social research (rev. ed.). — Sage, Newbury Park, CA.

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

— — & De Bruin, J.C. (1994). The ability of dogs to recognize and cross-match human odours. — Forensic Sci. Int. 69, p. 111-118.

Settle, R.H., Sommerville, B.A., McCormick, J. & Broom, D.M. (1994). Human scent matching using specially trained dogs. — Anim. Behav. 48, p. 1443-1448.

Sommerville, B.A., Green, M.A. & Gee, D.J. (1990). Using chromatography and a dog to identify some of the compounds in human sweat which are under genetic influence. — In: Chemical signals in vertebrates 5 (D.W. MacDonald, D. Muller-Schwarze & S.E. Natynczuk, eds). Oxford University Press, Oxford, p. 634-639.

Toner, B.S. & Miller D.I., Jr. (1993). Olfactory discrimination of individual human odors using experienced tracking police work dogs. — Anim. Behav. Consult. Newslttr. 10(4).

Wagenaar, W.A. & Veefkind, N. (1992). Comparison of one-person and many-person line-ups: a warning against unsafe practices. — In: Psychology and law: international perspectives (F. Lösel, D. Bender & Th. Bliesenser, eds). Walter de Gruyter, Berlin.

Walk, H.A. & Johns, E.E. (1984). Interference and facilitation in short-term memory for odors. — Percept. Psychophys. 36, p. 508-514.