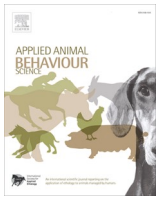
Applied Animal Behaviour Science 305 (2026) 107178

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

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



How does the behaviour of bio-detection dogs correlate with handler decision-making?

Z. Parr-Cortes [a,b], S. Morant [b], C.T. Müller [c], E. Drysdale [a], C. Guest [b], N.J. Rooney [a,b,*]

[a] Bristol Veterinary School, Bristol, BS40 5DU, United Kingdom
[b] Medical Detection Dogs, Milton Keynes, MK17 0NP, United Kingdom
[c] Cardiff School of Biosciences, Cardiff University, CF10 3AX, United Kingdom

ARTICLE INFO

Keywords:
Detection dog
Working dog
Dog training
Rating
Behaviour
Biomedical detection
Olfaction

ABSTRACT

During detection dog training, handlers must interpret dog behaviour to decide whether a target has been found. However, the consistency of handlers' decisions, and the behaviours they are based on, remain understudied. Eighty-four videos of 11 disease-detection dogs were shown to eight raters, blinded to search outcomes. They recorded: the dog's response using established decision categories; their certainty that the dog had identified a target (scale: 1–7); and the strength of the dog's response (scale: 1–5). Dog behaviour was also objectively analysed. Results showed moderate agreement between raters' decisions. We found many behaviours significantly associated with a target were also interpreted as indications by raters (e.g. sitting and time in contact with the sample port), while subtle behaviours like "head knock" were interpreted as interest or hesitation. Closer attention to such behaviours, and training of handler to avoid ambiguity in classification may improve the reliability of handlers' decisions, data recording and consequently dog training success.

1. Introduction

Dogs have been trained for the detection of many substances, such as narcotics (Jantorno et al., 2020; Lorenzo et al., 2003), explosives (Furton and Myers, 2001; Gazit et al., 2021; Lazarowski and Dorman, 2014; Oxley and Waggoner, 2009) and human diseases, including prostate cancer (Cornu et al., 2011; Guest et al., 2021), Parkinson's disease (Rooney et al., 2025) and COVID-19 (Grandjean et al., 2020; Guest et al., 2022; Sarkis et al., 2021). Medical Detection Dogs (MDD) is the UKs largest medical dog training charity. Here dogs are trained to sniff a series of odours and initially rewarded for showing interest in the target odour. A behavioural response to the target odour is then shaped and reinforced by the handler using positive reinforcement (Caldicott et al., 2023; Edwards et al., 2017; Mancini et al., 2015). Once the dog learns to identify the target odour and respond using a trained alert (also referred to as an indication), they interrogate each sample in the line (or around a carousel), deciding whether a target is present and demonstrating presence by performing an indication or absence by withholding the indication (Concha, 2023; Lazarowski et al., 2020).

Trained indications vary between dogs and can include lying down, sitting, standing, staring and nudging/pushing their nose to the sample port (Caldicott et al., 2023; Mancini et al., 2015). In addition, dogs will sometimes show a "response", to an odour of interest, such as a quick head turn, sometimes called a "head flick" or "check pace" (Mancini et al., 2015), or a change in body direction or position (Furton et al., 2010; Krichbaum et al., 2023). These spontaneous changes can help handlers decide whether a sample is a target, particularly if the dog does not perform their full trained indication (Mancini et al., 2015). Correctly identifying a target relies on the dog's ability to communicate that it has found a target via clear behavioural change and on the handler's ability to interpret the dog's behaviour (Mancini et al., 2015). Often, the handler must decide whether the dog's indication is sustained or strong enough to suggest that a target is present.

During dog training and testing, as well as recording alerts or indications versus no interest (i.e. no behavioural change) some organisations, like MDD, use intermediary categories of response, such as interest and hesitation to describe more ambiguous responses (Defence Science and Technology Laboratory (DSTL), 2022; Essler et al., 2020; Guest et al., 2020; Jezierski et al., 2015; Mancini et al., 2015; McCulloch et al., 2006; Sonoda et al., 2011). In published studies the definition of hesitation, for example, varies from "checking the sample twice or spending a long duration of time sniffing the sample without indicating" (Edwards et al., 2022; Essler et al., 2020) to "hesitating when sitting down, moving away from a sample but coming back to it" (Mancini

* Corresponding author at: Bristol Veterinary School, Bristol BS40 5DU, United Kingdom.
E-mail addresses: zc0342@my.bristol.ac.uk (Z. Parr-Cortes), Nicola.Rooney@Bristol.ac.uk (N.J. Rooney).

https://doi.org/10.1016/j.applanim.2026.107178
Received 6 April 2026; Received in revised form 25 June 2026; Accepted 7 September 2026
Available online 10 September 2026

et al., 2015). How categories are used is not always described, and to our knowledge, their reliability has not yet been empirically tested. It's important that organisations explore whether these multiple categories are useful, or whether they are perhaps used inconsistently. In this study, we used the established decision categories used at the time of the study by handlers at MDD as an example of a handler scoring system. We assessed how consistently the categories were used between raters; whether "interest" and "hesitation" are truly distinct categories, and we explored how the categorisation related to handlers' decision-making on whether the dog had found a target scent.

Aside from the dog's behaviour, other external factors such as the dog's performance on previous searches, events in the dog's or handler's life, changes in personnel or visitors and handler stress at the time of the search, can also influence handler decision-making and the dog's performance (Beebe et al., 2016; Caldicott et al., 2023; Defence Science and Technology Laboratory (DSTL), 2019; Jezierski et al., 2014; Zubedat et al., 2014). Familiarity with the dog has been shown to influence how handlers rate their dog's performance (Clark and Rooney, 2021). Hence we may expect that raters who are more familiar the dogs will be more consistent in their ratings, however, neither the consistency of decisions, nor the influence of rater's familiarity on their agreement with handlers, have been empirically tested.

The behaviours handlers use to determine whether a dog is showing an alert, and the consistency of these decisions across handlers, have not been empirically tested. We used pre-recorded videos and objective behavioural analysis, to explore how dog behaviour during detection searches affects raters' decisions on whether a target has been identified. The reliability of six established decision categories used by MDD was tested using inter-rater reliability analysis. We compared agreement to raters' familiarity with the dogs, and assessed whether greater familiarity with the dogs resulted in better agreement with the initial decision made. Since training and testing of detection dogs often involves a team of people, not just the dog handler, raters in this study included dog handlers and research staff, who were regularly present during the dogs' training and testing, and hence were routinely using the six decision categories. Finally, we compared dogs' behaviour while searching target and control samples to raters' decisions, to identify behaviours associated with a target and with the decision to record an indication, as well as behaviours associated with interest versus hesitation decisions. Together, these analyses provide insight into how handlers interpret dog behaviour during a detection task, and whether the existing classification is optimal or requires modification.

2. Methods

2.1. Ethical approval

Ethical approval for human participants was obtained from the University of Bristol Faculty of Health Science Research Ethics Committee (FREC) (Ref 12138) and for animal observation from the University of Bristol Animal Welfare Ethics Review Board (AWERB) (Ref UIN/22/008). Ethical approval for sample collection and use for individual studies was obtained and held by MDD and collaborating institutions and was not part of this study, as it did not alter dog training protocols.

2.2. Dogs and searches

Eleven bio-detection dogs, trained to find *in-vitro* remote samples of a range of diseases and odours, were filmed. They consisted of six neutered males and five neutered females, aged from 1 to 11 years, with between 1 month and 9 years of training as a bio-detection dog. Since it was the rater's decision-making we were studying, to explore the commonalities in behaviour across various dogs and odours, we deliberately selected a diverse array of dogs. There were seven Labrador Retrievers, two Cocker Spaniels, one Golden Retriever and one Hungarian Vizsla. Three dogs were trained to detect *Pseudomonas aeruginosa*, two prostate cancer (Guest et al., 2021), two *Escherichia coli*, two Parkinson's disease (Rooney et al., 2025) and two dogs were in early scent training, trained to detect an innocuous scent, e.g. a piece of Kong™ toy.

During routine training searches, individual stands were used to present a series of samples (3 or 4 per line-up). Each stand had a removable metal arm with a perforated plate on the end, behind which a sample jar was attached (Fig. 1a). The perforated plate acted as a sample cover and interface between the dog and the sample allowing the dog to sniff the sample without directly contacting it. Dogs were filmed searching for the target odour they were trained to detect: disease-positive sample of breath, sweat, urine or sebum (depending on the disease) or a piece of Kong™ toy (in early scent training) amongst control samples (disease-negative samples from individuals who are either healthy or had another condition, not the target disease) or non-target background odours e.g. water, empty jars, filter paper in early scent training. All target samples, control samples and blanks were placed in identical jars or containers. The number of target samples in the line-up was either zero, one or multiple. During each search, the handler stood behind a tinted glass screen positioned at the beginning of the line (closest to stand 1) so that the dog could not observe their body language (Fig. 1).

After the dog completed a first search or "pass" of the line-up of samples, the handler decided which of the following established decision categories best described the dog's behaviour at each sample. These were the standard categories routinely used by MDD at the time of video collection. The descriptions of the categories below are based on the raters' feedback, but definitions were not presented to the raters at the time as we were aiming to test their understanding of the categories:

- **Skipped (SKP)** – the dog did not approach the stand position as expected, but passed to the next position.
- **Not searched (NS)** – the dog did not search the sample as they didn't reach it due to an indication on an earlier sample).
- **No interest (NI)** – the dog searched the sample but showed no apparent change in behaviour
- **Interest (INT)** – the dog showed particular attention to the sample.
- **Hesitation (HES)** – the dog did not show a full indication, but lingered on the sample.
- **Indication (IND)** – the dog showed its trained alert while attending to the sample.

If an indication decision was made and the presence of a target was confirmed, the dog was rewarded (with Royal Canin Snacks Educ® treats), and the search ended. If there was no target present (false alert), the dog was not rewarded, and either continued to the next stand or the search was ended by the handler calling the dog away from the line. When the line-up consisted of only control samples (blank line), the dog was rewarded for not indicating at any stand and moving away or "coming off" the line after searching all samples. If the handler was uncertain if the dog had indicated or fully interrogated a sample, they would direct it to repeat the search i.e. perform a second "pass". A decision category was recorded for each sample in the line for every pass, regardless of whether the handler subsequently decided to repeat the line-up.

2.3. Video collection and rating

Over 3 months, a total of 27 h of video were collected from 55 bio-detection training sessions. Videos were recorded on a GoPro Hero 4 (GoPro, Inc., San Mateo, CA, USA) mounted on a stand near the location where the handler was positioned at the start of the line-up (Fig. 1b). A selection of 13 training session videos was pseudo-randomly selected, to include at least one session for each dog (eleven) and each handler filmed (four). Each video was edited into 13–35 video clips, each 5–20 s long and showing a dog searching a line of samples once. Each clip

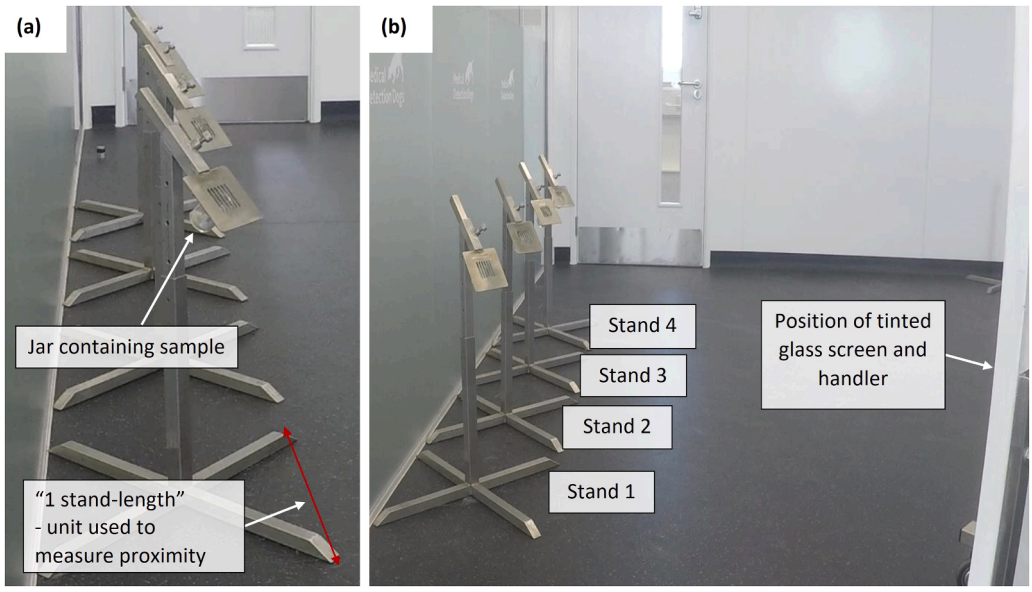


Fig. 1. (a) Stand line-up showing the location of a sample jar attached to the metal arm of stand 1. Jars were secured behind the metal perforated plate of each stand, through which the dog could sniff each sample. The red arrow indicates the length of 1 stand, used as a unit of measure for proximity. (b) The layout of the stand line-up, showing four stands labeled 1–4 and the position of the tinted glass screen behind which the handler stood.

started 5 s before the dog entered the frame and ended when the dog left the frame after moving away from the stands or after indicating towards a sample, but before any reward was given, so that a rater watching the clip would remain blind to the handler's response. Video clips with obvious cues of a correct or incorrect indication (e.g. handler intervention) were excluded to ensure the rater remained blind to the search outcome. The audio was removed from all clips to avoid auditory cues such as the handler speaking or the sound of the clicker (used as a secondary reward) revealing the outcome of the search and unblinding the rater. A total of 224 video clips were created. All were watched by the researcher (ZPC) blinded to the outcome of the search. For each stand in a clip, she recorded a decision category (SKP, NS, INT, HES or IND) along with a score for her certainty that the sample was a target (scale 1–7) and a score for how strong the dog's indication on the sample was (scale 1–5). Using these scores, a total of 84 video clips were selected. For each, a single, stand in the line (stand 1–4) was chosen for raters to observe and score. Stand selection aimed for approximately equal numbers of each: decision category, certainty score, indication strength score, dog, handler, disease type, original handler decision and position of the target in the line (stand 1–4). The clips included first, (N = 61) second (N = 21) or third (N = 2) "pass", i.e. the dog may be interrogating the samples for the first time, or it may be a repeat search of that line-up of samples

Eight dog handlers and research staff (with prior dog handling experience or currently routinely observing the dogs in this study during training and/or testing) working for MDD were recruited via announcements at team meetings. Volunteers were provided with an information sheet and consent form to review before deciding whether to participate. On the day of the study, raters were given the opportunity to ask any questions before signing the consent form and informed of their right to withdraw at any time.

For seven raters, the task was completed at their place of work (MDD training centre), with the researcher (ZPC) present. Before starting, the researcher read out the on-screen instructions and provided an opportunity for them to ask any questions. The task was presented on a computer, and raters completed questions at their own pace. The researcher remained in the room throughout the task in case there were any technical issues, and to prevent raters conferring, but did not comment on the content of the questions or video clips hence avoiding influencing raters' answers. For one rater who could not attend in person, the task was completed over video call with the researcher sharing their screen and controlling the video play/replay button at the rater's instruction, while the rater completed the questions on their computer. Each rater was presented with a Microsoft Form containing the 84 selected video clips of bio-detection dogs performing a single search on a sample line-up (3 or 4 stands). For each video clip, they were asked to score the dog's response for a single (predetermined) stand in the line. The stands were numbered from closest to the camera (1) to furthest away (4) (Fig. 1b). The stand the rater needed to observe was displayed above the video clip. Since some were very brief and intense concentration was required, raters were permitted to watch each video clip a maximum of twice and were advised to watch the entire video clip before answering any questions. They were informed that the handler in the video clip may or may not be blinded to the correct target location and that each search may have zero, one or multiple targets in the line. They were reminded that the task aimed to understand how handlers interpret dogs' behaviour and that there were no correct or incorrect answers. They were asked to base their decisions only on the dog's behaviour in the video clip and not on how it was edited or what they thought might happen once the clip ended (see Supplementary Materials for full instructions),. The task took approximately one hour to complete and included two 5-minute screen breaks (after 20 and 40 min). After completion, raters were asked not to discuss the video clips, questions, or their answers with the other raters.

For each video clip, raters answered the following three questions for the specified stand position. All raters were familiar with the decision categories used during routine dog training. Because the purpose of the study was to understand how raters interpreted these commonly used categories and assess whether they were using them consistently, full definitions for each category were not given with the questions.

1) Based on the behaviour shown in the video, what decision would you record for [stand number] using the categories below?
   Possible answers: SKP, NS, NI, INT, HES or IND
2) Based on the dog's behaviour, how sure are you that this sample is a target?
   Possible answers: Scale 1 (Absolutely certain it is not a target) to 7 (Absolutely certain it is a target) or N/A (sample not interrogated).
3) How strong is the dog's response/indication on this sample (i.e. response/indication that it is a target)?

Possible answers: Scale 1 (Very weak) to 5 (Very strong) or N/A (No response/indication).

After answering these questions for all 84 videos, raters were shown photos of the eleven dogs and asked how familiar they were with each by answering four questions on their experience with each dog (see Supplementary Materials for all questions).

2.4. Behaviour analysis

The same 84 clips were analysed by the researcher (ZPC) who was blinded to the stand position raters had been asked to score, any target location, the handler's original decision and the raters' responses. Behaviour at all stands in the video clip was analysed, including the stands that were not scored by raters. Of the 84 video clips, 13 had three stands in the line-up and 71 had four stands (this variation is common practice during dog training). Together, this resulted in a total of 323 individual stands available for the dog to search across all video clips (13×3 stands $+ 71 \times 4$ stands). Of these, 48 stands were not approached by the dog during the video clip (e.g. the video clip ended before the dog reached the stand or the dog moved away from the line-up before approaching the stand). These stands were recorded as "no video" and excluded from analysis. Therefore, there were 275 stands the dog approached and dog behaviour was recorded. A behavioural ethogram of 34 behaviours (Table S1 in Supplementary Materials) was created based on previous behavioural observation studies of medical detection dogs (Wilson et al., 2019; Wilson et al., 2020) with the addition of two behaviours mentioned by MDD handlers as important when observing dogs during searches: "eye flick" and "head knock". These 34 behaviours encompassed eight categories: body direction, gaze direction, head orientation, postural state, distance from plate, location, activity and interaction with the stand. Behavioural Observation Research Interactive Software (BORIS) was used to analyse dog behaviour at each of the 275 stands.

The frequency and duration of each behaviour at each stand were exported to Microsoft Excel, and the data were explored. Four behaviours ("body directed away from stand", "approach handler", "paw stand" and "lick lips") were excluded due to infrequent observations (< 5% of all 275 observations). Of the remaining 30 behaviours, five were excluded since they were observed during all searches or were deemed to have no biological relevance to the research questions: "move towards stand", "locomotion", "body directed forwards", "not looking at stands or handler" and "unknown gaze direction". Additionally, "stationary" was excluded as it was composed of "sitting" and "standing", which were analysed separately. "Looks towards observed stand" was excluded since "head oriented towards stand" was deemed more reliably determined than gaze direction. The behaviours "continues to another stand" and "comes off line" were excluded as they were both encompassed in "moves away from a stand". "Stays at stand" was excluded since it was the inverse of "moves away from stand". Due to the similarities between "mouth plate" and "nudge plate", a new behaviour, "push plate", was created that combined the frequency of these two behaviours. The final 19 behaviours used in the statistical analysis included 14 state behaviours and five event behaviours (Table S2 in Supplementary Materials).

2.5. Statistical analysis

2.5.1. How do indication strength and certainty that a sample was a target relate to the decision category selected by raters?

From a total of 672 responses (84 video clips × 8 raters), there were 89 "N/A" responses (8–14 N/A responses per rater) for at least one of the two ratings (handlers' certainty that the sample was at target or dog's indication strength), these responses were omitted from this analysis. This resulted in the exclusion of all "SKP" and "NS" responses, since these were all associated with an "N/A" rating for certainty and/or indication strength. In addition, one "NI" response that was associated with an "N/A" rating for certainty the sample was a target was also excluded. The following statistical analyses were run on the remaining 583 responses. To assess the correlation between handlers' certainty that the sample was a target and indication strength, a Spearman's Rank Correlation was conducted (using the cor.test functions from the stats R core package ver. 4.1.3) (R Core Team, 2013). To assess the relationship between decision category, indication strength and certainty ratings, a Cumulative Link Mixed Model with Laplace approximation was run for each rating (using the clmm function of the ordinal R package ver. 2022.11–16) (Christensen, 2022). Rater ID was included as a random effect to account for individual differences between raters.

2.5.2. Rater agreement

To assess agreement between raters across all six categories (SKP, NS, NI, INT, HES and IND) and all 672 responses, Kappa statistics were computed using the irr R package (ver. 0.84.1) (Gamer et al., 2019). The Fleiss Kappa function (kappam.fleiss) was used for comparisons between more than two ratings and the Cohen Kappa function (kappa2) was used for comparisons between two ratings. Kappa statistics produce a value between 0 - s no agreement, and 1 - perfect agreement. We classified agreement categorisation as Landis and Koch (1977): poor ($k \leq 0.20$), fair ($k = 0.21–0.40$), moderate ($k = 0.41–0.60$), substantial ($k = 0.61–0.80$) or near perfect ($k \geq 0.81$). Rater agreement was assessed to answer the following research questions:

2.5.2.1. Do raters agree on decision categories?. Agreement across all eight raters was assessed overall and for each of the six decision categories (SKP, NS, NI, INT, HES and IND) using the Fleiss Kappa statistic.

2.5.2.2. Do the majority of raters agree with the handler's original decision?. The original handler's decision (recorded at the time of the search) was compared to the majority vote (category selected by ≥ 4 out of 8 raters) using the Cohen Kappa statistic. Of the 84 video clips, 7 had no majority decision (< 4 votes for one category) and were excluded from this analysis. A heat map of the confusion matrix showing the number of correct classifications (agreement) between the handler's original decision and the raters' majority decision was constructed (using the caret ver. 6.0–94 (Kuhn, 2008) and ggplot2 ver. 3.4.4 (Wickham, 2016) R packages).

2.5.2.3. Does greater familiarity with the dog increase raters' agreement with the handler's original decision?. Cohen's Kappa statistic was calculated for each rater, comparing their decision with the original handler's decision. A Spearman's rank correlation was used to compare individual raters' kappa statistics to their total familiarity score. The maximum possible familiarity score was 55 (5×11 dogs), with higher scores indicating greater familiarity with all dogs. Three raters were also handlers in some of the video clips, these raters are indicated by an asterisk in the results table.

2.5.3. Which behaviours were associated with the dog encountering a target?

Of the 275 stands where dog behaviour was analysed, the original handler's decision for 25 stands was "NS" or "SKP". Since samples at these stands were determined not to have been searched, any targets placed here may not have been interrogated by the dog and were excluded from this analysis. To assess if dog behaviour was affected by the sample type (target or control), binomial generalised linear mixed-effects models (GLMM) were used. Frequencies and durations of behaviours could not be modelled using a Poisson distribution due to an excess of zeros. Therefore, each variable was dichotomised; frequency behaviours as zero vs non-zero counts i.e. whether the behaviour occurred (1) or not (0) and duration behaviours as above or below the upper quartile (based on the distribution of data). For example, if the duration the dog's nose was in contact with the plate was > 0.47 s (upper quartile), it was scored as 1 and if the duration was ≤ 0.47 s, it

was scored 0. For some duration behaviours (e.g. looking at the handler), the upper quartile was equal to zero; therefore, if the duration of these behaviours was 0 s, it scored 0, otherwise it scored 1. Binary values were then used in a binomial GLMM to calculate the log odds ratio (log(OR)) of each behaviour occurring when the dog encountered a target sample or not (i.e. control sample). Behaviours more likely to occur when encountering a target had log(OR) > 0, and behaviours less likely to occur had log(OR) < 0. Dog ID was included as a random effect to control for individual dog differences.

2.5.4. Which behaviours were associated with the raters' decision to select the "indication" category?

From a total of 672 rater responses (84 video clips × 8 raters), there were 88 "NS" (n = 46) or "SKP" (n = 42) rater responses (8–14 responses per rater). Since the rater determined samples at these stands were not searched, these responses were excluded from this analysis. The remaining 584 responses (NI: n = 134, INT: n = 122, HES: n = 120, IND: n = 208), were used to assess which behaviours were associated with the raters selecting the IND category compared to each of the other three categories: NI, INT and HES, in three separate binomial GLMMs for all 19 behaviours recorded (i.e. IND vs NI, IND vs INT and IND over HES). Rater ID was included as a random effect to control for individual rater differences.

2.5.5. Are interest and hesitation categories truly distinct?

We assessed both the inter-rater reliability of raters selecting these categories and the objective differences in behaviours associated with raters selecting HES compared to INT. To do so, we asked the following two questions:

2.5.5.1. Which behaviours were associated with the hesitation compared to the interest decision category selected by raters?. Binomial GLMMs were used to calculate the log(OR) of raters selecting HES compared to INT for all 19 behaviours (using the glmer function of the lme4 R package ver. 1.1–29) (Bates et al., 2015). Rater ID was included as a random effect to control for individual rater differences.

All analyses in this study used R Studio Version 2023.06.2 + 561. Plots were produced using the ggplot2 (ver. 3.4.4) (Wickham, 2016), ggstats (ver. 0.5.1) (Larmarange, 2023) and ggpubr (Kassambara, 2023) R packages.

2.5.5.2. How does combining interest and hesitation categories affect rater agreement levels?. INT and HES decisions were combined into a single "INT/HES" category, and the Fleiss Kappa was calculated to assess agreement across all eight raters. The overall agreement level and individual agreement levels for each of the five new decision categories (SKP, NS, NI, INT/HES and IND) are reported.

3. Results

3.1. How do indication strength and rater certainty that a sample was a target relate to the decision category selected?

There was a strong and significant positive correlation (r = 0.84, p < 0.001) between indication strength and certainty scores for responses that contained a numerical rating (i.e. excluding N/A responses that were associated with NS and SKP responses). Cumulative Link Mixed Models showed raters scored significantly higher indication strength and were significantly more certain that a sample was a target when selecting IND > HES > INT > NI (p < 0.001). Pairwise comparisons found that the differences between the four categories were all significant for both indication strength (p < 0.001) and certainty (p < 0.001), with the smallest differences in scores between INT and HES categories.

3.2. Rater agreement

3.2.1. Do raters agree on decision categories?

Overall, there was moderate agreement between raters across all decision categories (Fleiss kappa = 0.57, z = 55.9, p < 0.001). Agreement was substantial for the IND and NI categories, moderate for the SKP, INT and HES categories and poor for the NS category (Table 1).

3.2.2. Do the majority of raters agree with the handler's original decision?

The agreement between the majority rater decision (category selected by ≥ 4 out of 8 raters) and the original handler decision (recorded at the time of the search) for the 77 video clips with a clear majority was substantial (k = 0.75, z = 12.9, p < 0.001). A heat map of the confusion matrix (Fig. 2) shows the highest frequency of agreement (24) was observed for the IND category, and the most frequent mismatches were between INT and HES. Of the 13 HES decisions made by the majority of raters, five were classified as INT (38.5%) by the original handler. Of the seven videos with no clear majority, four had an equal number of HES and IND ratings (four votes each), three were recorded as IND, and the fourth was recorded as INT by the original handler. Another two videos with no majority decision had an equal number of NS and SKP ratings, one of which was recorded as SKP and the other NI by the original handler. The final video that had no majority decision had an equal number of INT and HES ratings and was recorded as INT by the original handler.

3.2.3. Does greater familiarity with the dogs increase raters' agreement with the handler's original decision?

When looking at the agreement between individual raters and the original handler decision, the level of agreement varied from moderate to substantial (k = 0.49–0.66, p < 0.001) (Table S3 in Supplementary Materials). We found no correlation between individual Cohen Kappa statistics (raters' agreement with the original handler's decision) and familiarity scores (raters' familiarity with the dogs in the videos) (r = 0.12, p = 0.84).

3.3. Which behaviours were associated with the dog encountering a target?

We found that 14 of the 19 behaviours analysed were significantly more likely to be observed when

the dog encountered a target (log(OR) > 0, p < 0.05), one was significantly less likely to occur (log(OR) < 0, p < 0.05), and four were not significantly associated with the presence of a target (Fig. 3). Sitting (duration > 0 s), licking the plate (frequency > 0), orienting head towards the stand (duration > 1.61 s), contact with the plate (duration > 0.47 s) and body directed towards the stand (duration > 0.2 s) were over twice as likely to be observed when a target sample was encountered versus a control sample (log(OR) > 2) (p < 0.001). Moving away from the stand (frequency > 0) was almost four times *less* likely to occur when encountering a target versus a control sample (log(OR) = –3.9, p < 0.001) (see Table S4 in Supplementary Materials for full results).

3.4. Which behaviours were associated with the raters' decision to select the "indication" category?

Of the 19 behaviours analysed, increases in the frequency or duration of 10 behaviours: "lick plate", "push plate", "> 1 nose-length from plate", "head oriented away from stand", "wagging tail", "< 1

nose-length from plate", "body directed towards stand", "head oriented towards stand", "time at stand", and "contact with plate" (p < 0.0010) resulted in raters being significantly more likely to select IND compared to NI, INT and HES (Fig. 4). A longer duration of two further behaviours: "sitting" and "looking at handler" and increased frequency of "eye flick" prompted raters to select IND compared to INT and HES categories (p < 0.001) (Fig. 4b and c). This suggests these 13

Table 1
Results of Fleiss Kappa for individual categories across all eight raters.

Decision Category	Kappa	z value	p value	Classification
Indication (IND)	0.80	38.9	< 0.001	Substantial
No Interest (NI)	0.71	34.4	< 0.001	Substantial
Skipped (SKP)	0.47	22.8	< 0.001	Moderate
Interest (INT)	0.41	19.8	< 0.001	Moderate
Hesitation (HES)	0.41	19.7	< 0.001	Moderate
Not Searched (NS)	0.29	14.2	< 0.001	Fair
Overall agreement across all categories: Fleiss kappa = 0.57, z = 55.9, p < 0.001				

Overall agreement across all categories: Fleiss kappa = 0.57, z = 55.9, p < 0.001

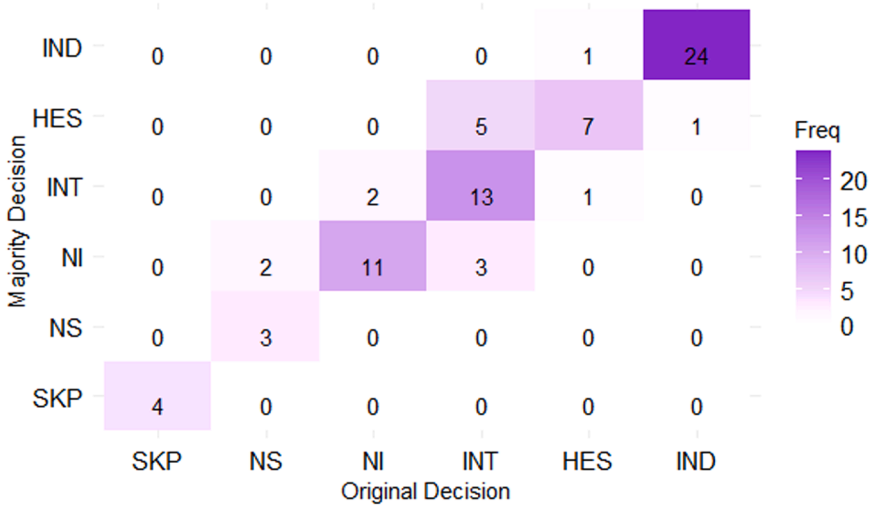


Fig. 2. Heat map showing agreement between the majority rater decision (category selected by ≥ 4/8 raters) and the original decision (handler decision at the time of the search) based on 77 videos where there was a majority decision. Purple shading and values represents the frequency (freq) of occurrences for each combination of the original decision and majority decision. Darker shades indicate higher frequencies and lighter shades indicate lower frequencies. SKP = Skipped, NS = Not Searched, NI = No Interest, INT = Interest, HES = Hesitation, IND = Indication.

behaviours, were consistently perceived as an indication by raters. The duration of "body directed backwards" was not significantly associated with raters selecting IND compared to INT or HES, suggesting this behaviour was perceived as either interest, hesitation or an indication.

A longer duration of "standing" was associated with a significantly higher likelihood of raters selecting IND versus NI and INT, but no significant difference between IND and HES decisions, suggesting this behaviour was perceived as either hesitation or an indication. An increase in "head knock" behaviour was associated with a raters being significantly more likely to select IND compared to NI ($p < 0.001$) (Fig. 4a), but *less* likely to select IND compared to INT ($p = 0.004$) and HES ($p = 0.006$)

(Fig. 4b and c), suggesting this behaviour was perceived as either interest or hesitation (see Table S5 in Supplementary Materials for full results).

3.5. Are interest and hesitation categories truly distinct?

3.5.1. Which behaviours are associated with raters' decision to select the "hesitation" category compared to the "interest" category?

Raters were significantly more likely to select HES compared to INT when there was a higher frequency or duration of 15 out of 19 behaviours ($p < 0.05$). Only the occurrence of "head knock", "looking at another stand", "lick plate" and "sitting" was not significantly different when raters selected HES or INT (Fig. 5) (see Table S6 in Supplementary Materials for full results).

3.5.2. How does combining the interest and hesitation categories affect agreement levels?

Since there was overlap in the confusion matrix for INT and HES categories (Fig. 2), we tested whether combining INT and HES into one category improved rater agreement across all categories. After combining HES and INT, the overall Fleiss Kappa across all categories increased from k = 0.57 (Table 1) to k = 0.67 (Table 2), with the classification increasing from moderate to substantial. Agreement for individual INT and HES categories also increased from moderate for INT and HES independently (both k = 0.41) to substantial for the combined INT/HES category (k = 0.66).

Fig. 6.1. Overview of experimental design. Coloured boxes and letters indicate the odour type collected from volunteers (phase 1) or presented to dogs (phase 2): "S" (orange) = stress odour, "R" (blue) = relax odour, "B" (grey) = blank cloth odour, "X" = no odour (closed jar during training). The order of odour exposure in phase 2 was counterbalanced across all 18 dogs: Order 1 = stress odour in session 2 and relax

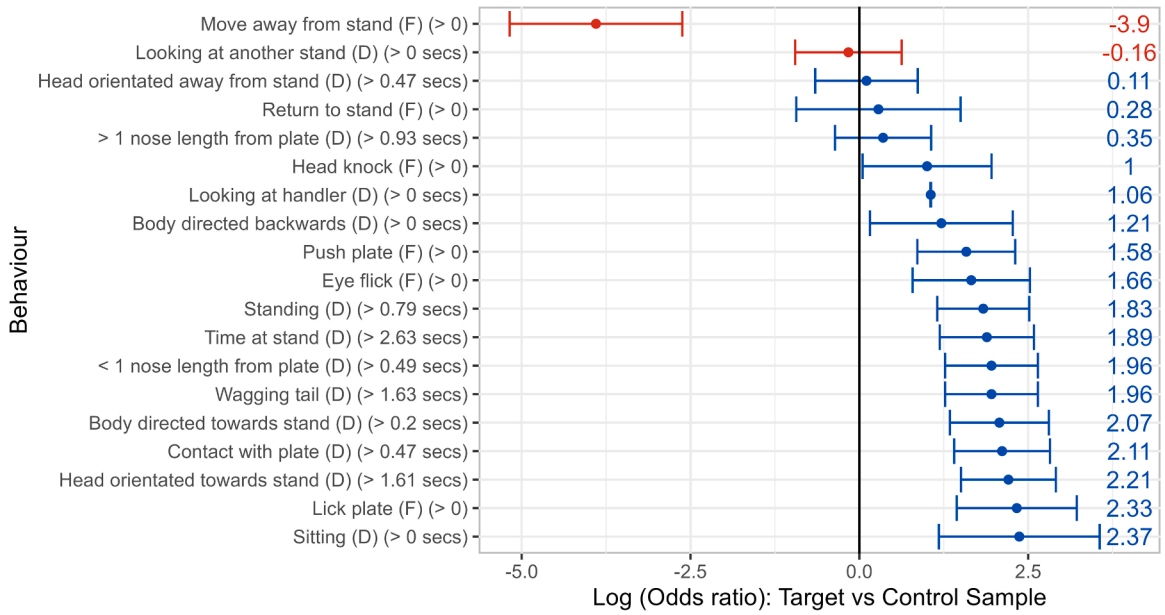


Fig. 3. Log odds ratio (log(OR)) for associations between each behaviour and whether the sample was target (versus control). Thresholds for the presence of the behaviour were set to > 0 counts for frequency behaviours (F) and > the upper quartile (secs) for duration behaviours (D). Behaviours more likely to occur when encountering a target (log(OR) > 0) are shown in blue, and behaviours less likely to occur when encountering a target (log(OR) < 0) are shown in red. The solid black vertical line indicates log(OR) = 0. Log(OR) are shown on the right. Error bars indicate 95% confidence intervals.

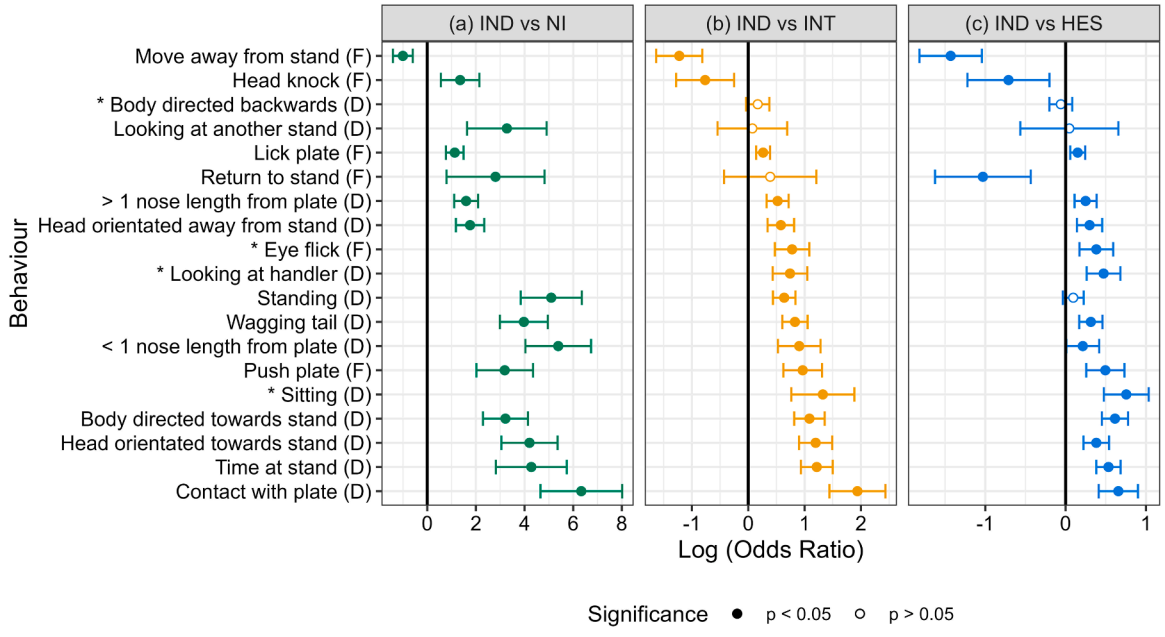


Fig. 4. Log odds ratios (log (OR)) for associations between each behaviour and raters' selection of the indication (IND) category compared to (a) No interest (NI), (b) Interest (INT) and (c) Hesitation (HES). Frequency behaviours are indicated by (F), and duration behaviours are indicated by (D). Behaviours observed more often/for longer when IND was selected have log(OR) > 0 and behaviours observed less often/for less time when IND was selected have log(OR) < 0. The solid black vertical lines indicate log(OR) = 0. Error bars indicate 95% confidence intervals. Solid circles indicate significance $p < 0.05$, and open circles indicate $p > 0.05$. Behaviours indicated with an asterisk (*) were never observed when NI was selected, therefore log(OR)s for IND vs NI for these behaviours are infinite and not plotted here.

odour in session 3, Order 2 = relax odour in session 2 and stress odour in session 3.**Figure 5.6.** Plot showing log odds ratios (log (OR)) associated with the selection of the hesitation category (HES) compared to the interest category (INT) for each behaviour. Frequency behaviours are indicated by (F), and duration behaviours are indicated by (D). Behaviours observed more often/for longer when HES was selected have log (OR) > 0, and behaviours observed less often/for less time when HES was selected have log(OR) < 0. The solid black vertical line indicates log (OR) = 0. Error bars indicate 95% confidence intervals. Solid circles indicate $p < 0.05$, and open circles indicate $p > 0.05$.

4. Discussion

When using their established decision categories to rate dogs'

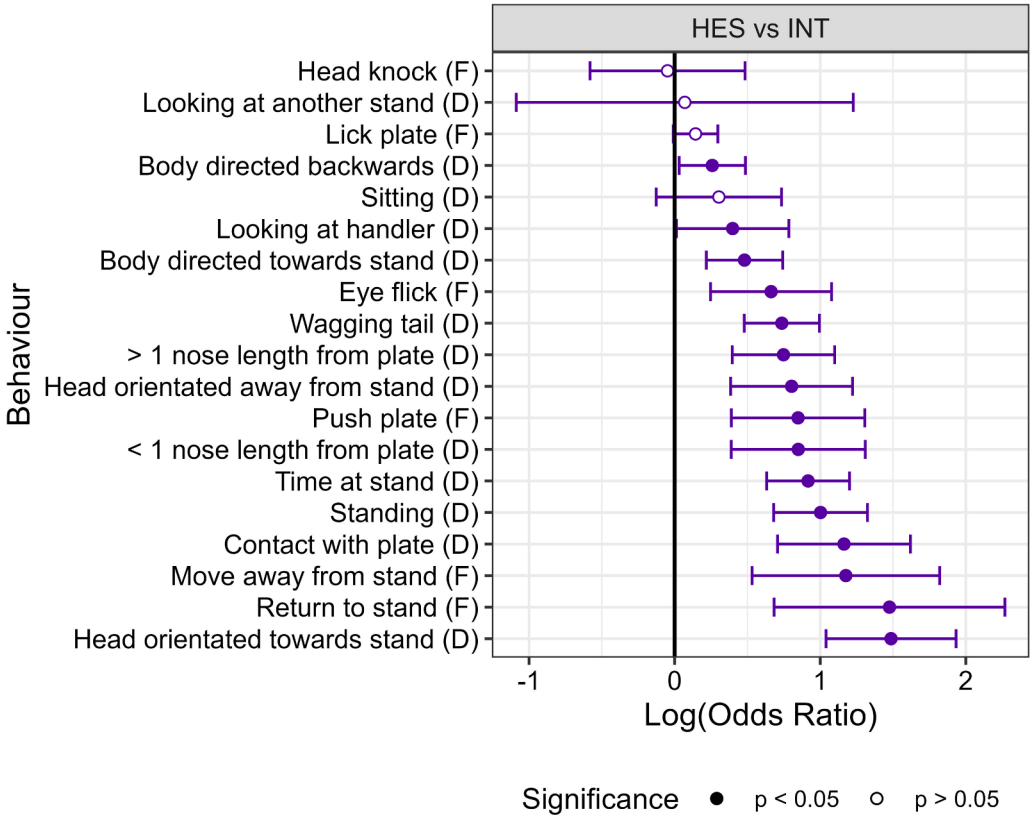


Fig. 5. Log odds ratios (log (OR)) associated with the selection of the hesitation category (HES) compared to the interest category (INT) for each behaviour. Frequency behaviours are indicated by (F), and duration behaviours are indicated by (D). Behaviours observed more often/for longer when HES was selected have log (OR) > 0, and behaviours observed less often/for less time when HES was selected have log(OR) < 0. The solid black vertical line indicates log(OR) = 0. Error bars indicate 95% confidence intervals. Solid circles indicate $p < 0.05$, and open circles indicate $p > 0.05$.

Table 2
Results of Fleiss Kappa for individual categories across all eight raters after combining interest and hesitation categories into one category (INT/HES).

Decision Category	Kappa	z value	p value	Classification
Indication (IND)	0.80	38.9	< 0.001	Substantial
No Interest (NI)	0.71	34.4	< 0.001	Substantial
Interest/Hesitation (INT/HES)	0.66	32.0	< 0.001	Substantial
Skipped (SKP)	0.47	22.8	< 0.001	Moderate
Not Searched (NS)	0.29	14.2	< 0.001	Fair
Overall agreement across all categories: k = 0.67 (z = 55.4, p < 0.001)				

Overall agreement across all categories: k = 0.67 (z = 55.4, p < 0.001)

behaviour during bio-detection searches, we found overall moderate agreement across eight raters. The "indication" and "no interest" categories had the highest agreement levels and "not searched" the lowest. There was a strong positive correlation between indication strength, raters' certainty that the sample was a target and the decision to select IND > HES > INT > NI; suggesting that raters were most certain a target was present when selecting IND and that this was associated with their perception of the strongest behavioural response. The ratings of indication strength and certainty that the sample was a target were significantly different between all categories, with the smallest differences between the INT and HES categories. Using objectively recorded dog behaviour, we found that most, but not all, behaviours that were significantly more likely to occur when a target was encountered were also associated with raters' decision to record an indication. Because "not searched" and "skipped" responses were consistently associated with "N/A" ratings for certainty the sample was a target and indication strength, these categories could not be analysed for correlation between ratings. Neither could they be analysed for their association with dog behaviour since the dog spent little to no time at the stand when these categories were selected.

4.1. Rater agreement

The eight raters showed moderate agreement across all decision categories, with the greatest agreement for IND and NI categories (Table 1). This is unsurprising since these categories were associated

with the strongest and weakest indication strength ratings and the most and least certainty that the sample was a target, respectively. This indicates that there was a consensus among raters on the behaviours that clearly signalled the presence (IND) and absence (NI) of a target. This is likely because raters hold clearer a-priori representations of the difference between "indication" and "no interest" since they will have a concept of a "find" versus "no find". The comparison between the raters' majority decision and the original handler's decision also demonstrated that there was little overlap between IND and other categories, again indicating high agreement levels (Fig. 2). This is encouraging since recognising and recording an indication is the most important decision in determining whether a target has been detected. Similarly, an NI decision made by the original handler at the time of the search was rarely confused with other categories by blinded raters (Fig. 2); only 2 out of 13 decisions (15.4%) confused for INT, indicating high agreement. Interestingly, there was slightly more confusion in the opposite direction, i.e. there were more instances when blinded raters selected NI when the original decision had recorded NS (12.5%) or INT (18.8%). This could be due to the differences in vantage point of raters observing the video clips (and raters' ability to replay the video clip), versus handlers deciding in real-time with less time to consider their decision. In these instances, recording NS or INT could prompt the handler to direct the dog to interrogate the sample again before determining if the dog truly showed no interest in the sample.

The poorest agreement between raters was for the "not searched" (NS) category ($k = 0.29$) (Table 1). This is surprising since we may expect this to be clear and easy for raters to agree on. One possibility for this finding is that determining whether a dog has adequately sniffed a sample can be difficult. This is especially true in the absence of audio, since the dog's nose can be close to the plate without the dog sniffing the sample. While we acknowledge that the lack of audio may have also limited the cues used by raters, it was necessary to ensure raters remained blinded to the correct response, which may otherwise have been apparent through audio cues. In contrast, there was moderate agreement for the "skipped" (SKP) category (Table 1), likely because it is more obvious to see when the dog does not get close enough to a stand to search the sample and instead moved passed it. It is important to note, that there were fewer NS and SKP responses ($n = 42$ and $n = 46$, respectively) than the other four categories ($n = 120–208$), which means the weighting of one disagreement would be numerically greater than for the other categories. We also found that these responses were associated with fewer behaviours being displayed, since the dog not spending as much time at the stand or was not getting close enough to the port for behaviours to be recorded. Similarly, these responses often resulted in "N/A" ratings from raters, making it impossible to analyse raters' interpretation of these categories. Since our focus was to examine dog behaviour and indication strength while interrogating samples, we included fewer videos where the sample was not searched. To investigate the reliability of NS and SKP categories further, future studies should include more videos where these categories are recorded, as well as rater questions and a behaviour ethogram more specifically designed to look for behaviours that are less sample oriented, or that occurred further from the sample port or stand. It could also include analysis to determine whether NS and SKP are considered truly distinct categories, or if they should be combined, similar to that performed for the INT and HES categories in the current study.

Although the agreement between raters and the original handler varied across raters from moderate to substantial (Table S3 in Supplementary Materials), we found no significant correlation between raters' familiarity with the dogs and their agreement with the original handler's decision. Although not empirically measured in this study, it's possible that other factors, such as level of experience and/or training, may have influenced individual agreement levels. There is evidence, however, that the length of time or level of experience working or living with dogs does not result in an increased ability to interpret dog behaviour correctly and that this varies between individuals regardless of their level of experience (Tami and Gallagher, 2009).

4.2. Behaviours associated with target samples and/or indication decisions

Attempts to explore the relationship between rater decisions and combinations of dog behaviours resulted in unstable models due to high correlations between behaviours. Therefore, it was not possible to model a combination of behaviours that best predict a target or an indication, and instead each behaviour was analysed sequentially. Although we report p-values, for these exploratory analyses more can be gained from examining effect sizes and overall patterns of behaviours. This behavioural analysis showed that 11 out of 19 behaviours measured were significantly associated with both the dog encountering a target sample and the raters' selection of IND compared to NI, INT and HES categories (Fig. 6). Together, these 11 behaviours indicated the presence of a target and were reliably recognised by raters as an indication. Many of these behaviours are consistent with trained alerts. For example, sitting, pushing or licking the plate, time in contact with or < 1 length from the plate and overall time at the stand. These behaviours are indicative of the dog spending longer in front of, near and interacting with the sample and are behaviours that the raters correctly identified as indicating the presence of a target. These results also suggest that a sit indication was the best predictor of a target sample, since it was over twice as likely to occur when a target sample was encountered compared to a control sample (Fig. 3). Since sitting was also associated with raters selecting IND compared to INT and HES and was never observed when raters selected NI, it was an easily recognisable and reliable indicator of the dog identifying a target sample.

Of the 14 behaviours associated with a target, three were not consistently associated with raters' decision to record an indication. These were standing, "head knock", and orienting their body backwards (toward the start of the line-up) (Fig. 6). Although standing for longer than 0.79 s was a significant predictor of a target sample, the duration of standing was not significantly associated with raters' decision to select IND compared to HES. This suggests that dogs spend a similar amount of time standing in front of a sample when perceived to "hesitate" as when they were perceived to indicate. This is in line with findings by Essler et al. (2020) who found that dogs that used a stand and stare indication showed more hesitation than dogs with a sitting indication. However, it's also possible that this overlap was due standing being recorded any time the dog stopped moving and could occur even if a clear indication was not performed e.g. the dog simply spending a longer sniffing a sample.

One of the "stimulus-response" behaviours, "head knock", also sometimes described as a "head flick" or "check pace" (Mancini et al., 2015) was significantly associated with the dog encountering a target. However, it was significantly more frequent when raters recorded INT and HES compared to IND. One possible explanation is that although a "head knock" was displayed when dogs recognised a target odour, but unless the dog showed other behaviours suggestive of an indication, this behaviour alone was not recognised by raters as identification of a target. The nuances of these subtle behaviours and their influence on handler decision-making have often been discussed in the literature, with calls for further analysis and insight into ways to support handler decision-making (Caldicott et al., 2023; Concha, 2023; Mancini et al., 2015). The results of the current study suggest that "head knock" may be an important subtle behaviour to which handlers may need to pay more attention. Misinterpretation or a lack of sensitivity to subtle behaviours can increase false negatives recorded during scent detection. Wasser et al. (2004) report that the most common reason for dogs to "miss" planted samples of scat was handlers moving their dogs away from a sample without realising the dog had found it. Therefore, being able to recognise subtle behaviours, including spontaneous stimulus responses, can help the handler decide whether a sample is a target (Mancini et al., 2015).

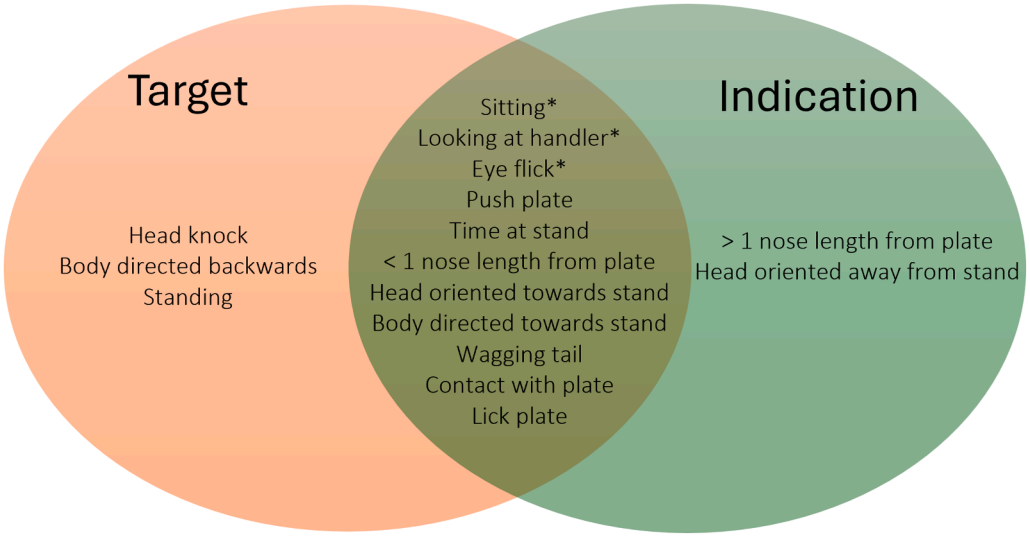


Fig. 6. Venn diagram showing overlap in the behaviours associated with the dog encountering a target sample (orange) and raters' decision to record an indication (IND) (green), based on results of generalised linear mixed models (GLMM). * Behaviours never observed when no interest (NI) decision recorded.

Two additional behaviours were significantly associated with the raters' selection of IND, but were not predictors of a target sample (Fig. 6). These were spending longer > 1 nose-length from the plate and orienting their head away from the stand. The three proximity behaviours were all associated with raters selecting IND compared to NI, INT and HES. Since the sum of these three behaviours equals the time spent at the stand, this is likely due to the dog spending longer at the stand (regardless of how close their nose was to the plate), which may have been interpreted as an indication. However, only time in contact and time < 1 nose-length from the plate were associated with the dog encountering a target. This suggests there were instances when dogs spent longer with their nose > 1 nose-length from the plate, which were interpreted as an indication by raters (possibly due to the presence of other behaviours), but these were not at target samples, i.e. false alerts.

There is evidence that dogs sniff for significantly longer when a target is present, even when they do not perform a final indication or "miss" the target (false negative) (Concha et al., 2014). Further consideration and processing of the odour may occur when a target is present, but other factors (such as uncertainty or lack of confidence) may prevent the dog from committing to a final indication. The length of time dogs spend with their nose near the sample may help handlers decide whether the dog has correctly ignored a control sample or whether a target has been missed. Assuming the dog has adequately searched the sample, spending longer with their nose close to or in contact with the sample port (even if they do not perform an indication), could be a prompt for handlers to direct the dog to check the sample again to ensure it is not a missed target.

Essler et al. (2020), found that dogs with a sit response spent less time sniffing a sample and had a higher false alert rate than dogs with a stand and stare response who generally spent longer sniffing a sample. The authors suggested this is because dogs using a sit indication moved their head away from the sample in order to perform their sit response (therefore spending less time sniffing a sample), whereas dogs using a standing indication were able to keep their nose closer to the stand when performing their indication (Essler et al., 2020). In theory, this allows dogs with a standing indication more time to consider their decision while holding their indication than dogs that use a sit indication, which may make quicker decisions based on less olfactory information (Caldicott et al., 2023; Essler et al., 2020). Therefore, it is important for handlers to ensure the dog has searched the sample for long enough before deciding if an indication is a correct or false alert and if need be, repeat the search if the indication seems premature. Since this study concluded, MDD have adjusted their training protocols and increased the time the dog must spend with its nose close to stand when performing an indication. This adjustment was made to avoid dogs turning their heads away from the stand prematurely and reducing the time spent sniffing the sample.

4.3. Behaviours associated with hesitation and interest decisions

The behaviour "body directed backwards", which described when the front of the dog's body was facing the direction it came from, sometimes reflected instances when the dog moved to the next stand before returning to the observed stand, resulting in the dog's body changing orientation. However, while "body directed backwards" was significantly associated with the dog encountering a target, the behaviour "return to stand" was not. This suggests there were instances when dogs turned their body to face the direction they came from without moving on, and these were often associated with a target sample. In addition, the dog spending longer with its body directed backwards was associated with raters' selection of either HES or IND, whereas the dog returning to the stand after moving on was associated with the raters' selection of HES more often than IND (Fig. 4c). These results suggest that both behaviours were associated with perceived hesitancy, with return to stand being more so. In contrast, moving away from the stand (and not returning) was the only significant behavioural predictor of a sample *not* being a target (i.e. a control sample) (Fig. 3) and was also associated with raters' selection of NI, INT and HES compared to IND (Fig. 4), suggesting it was a reliable and easily recognisable indicator of the absence of a target sample.

Agreement between raters on the intermediate categories (INT and HES) was poorer than IND and NI, but still moderate (Table 1). In addition, when comparing the majority rater decision to the original handler decision, the greatest confusion was between INT and HES (Fig. 2). This is unsurprising since these categories are often not well-defined. However, differences in perceived indication strength and certainty ratings between INT and HES were significant, indicating that raters considered these two categories as distinct. There were also significant differences for 15 out of 19 behaviours between clips rated as INT compared to HES, suggesting this distinction was based on objective behaviours (Fig. 5). These intermediate categories may serve as a useful tool in the decision-making process of individual handlers before coming to a final yes/no decision as to whether a target is present or whether the dog should search the sample again.

4.4. Improving inter-rater reliability and handlers' assessments of dog behaviour

There is evidence that benchmarking of rating scales can reduce ambiguity and ambivalence in scoring working dog performance (Clark and Rooney, 2021). In addition, standardised training such as frame-of-reference training and the provision of descriptions for each point on a scale have also been shown to improve rater accuracy and inter-rater reliability both independently and even more so when combined (Clark and Rooney, 2021; Melchers et al., 2011; Noonan and Sulsky, 2001; Schlientz et al., 2009). The inclusion of common descriptions and the provision of examples could be useful methods, not only to improve the inter-rater reliability of existing staff, but also as an ongoing measure to ensure both new and existing staff have a common understanding of the definitions of each category. Such training should ideally be accompanied by regular inter-rater reliability checks to ensure adequate agreement levels are maintained. In combination with these measures, more studies like the one described here may aid in identifying and recognising subtle differences in behaviour that may guide handlers' decisions.

While these findings are particularly relevant for bio-detection dogs and MDD, the principles of combining objective behaviour analysis with handler interpretation is applicable to a larger range of dog working roles. Here for standardisation, we analysed and rated the dog's behaviour and at single individual stands and included videos of 1st, 2nd and 3rd passes to ensure a diverse range of responses and behaviours. In a real-life training and testing scenarios, handlers' decisions are likely influenced by how the dog behaves in previous stands and passes. This may add accuracy, but can also introduce potential biases. Similarly, auditory signals in live working may give additional clues and hence increase classification reliability.

There are several other factors in the experimental paradigm that may have influenced agreement levels relative to real life, such as: raters' ability to re-play videos, the inclusion of dogs trained on different odours and the mix of handlers and research staff. These choices aimed for a diverse and representative sample of dogs, behaviours, observers and decisions and allowed us to identify common behavioural cues and assess agreement across dogs and tasks, rather than characterising the behaviour of individual dogs or handler-dog dyads. With more dogs and handlers, the next phase of research should investigate whether agreement differs between handler's rating their own dogs, handlers rating unfamiliar dogs, and non-handler observers, as well as examining how knowledge of the context of the full search (not just the individual stand) and the individual dog impact rater decision making.

5. Conclusion

This study explored the relationship between dog behaviour and handler decisions during a bio-detection task. We found moderate agreement across all eight raters and measurable differences in the behaviour of eleven bio-detection dogs when encountering a target odour compared to control odours. Most, but not all, behaviours associated with a target were also associated with the raters' decision to record an indication, suggesting that although raters did not always agree, they were generally effective at recognising when a dog had identified a target sample. However, we found a few behaviours that were associated with a target sample that were not perceived as an indication by raters. Similarly, some behaviours associated with an indication decision were not associated with the dog encountering a target. Raters used their understanding of the decision categories, acquired within their regular work, which highlighted some ambiguity in understanding. Our findings show the importance of clear, well-defined decision categories when assessing working dog performance and the need for standardised training to maintain good inter-rater reliability. They provide insight into how handlers perceive some more ambiguous behaviours as signs of uncertainty or hesitation, which in turn affects their decisions. More broadly, they demonstrate the complex relationship between handler and dog and the value of understanding the nuances of dog behaviour during detection tasks. Whilst ongoing work to automate decision-making and recording in detection dog work (e.g se Kulgod et al., 2026) may help to overcome these challenges, deep understanding and analysis of dog behaviour will likely remain critical to successful operations. In light of these results the organisation studied (MDD) have implemented additional benchmarking of decision categories and improvements to both dog and handler training. We suggest that rigorous analysis and validation of recording processes may be similarly valuable to other dog training institutions.

Author contributions

ZPC, NJR and CG designed and planned the study. ZPC collected the data. ZPC and ED created the behaviour ethogram. All video analyses, data analyses and interpretation included in this study were conducted by ZPC. SM advised and provided guidance on the statistical methods. ZPC drafted the manuscript, with input and supervision from NJR and CM and with editing and review by all authors.

Consent to participate

All participants provided informed consent and were reminded of their right to withdraw at any time without explanation.

Consent for publication

Informed consent for publication of anonymised data was provided by the participants.

Funding statement

This research project was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) as part of the South West Biosciences Doctoral Training Partnership (SWBio DTP) studentship awarded to ZPC. This studentship is a collaboration between the University of Bristol, Cardiff University and is part of a Collaborative Awards in Science and Engineering (CASE) studentship and Medical Detection Dogs is the CASE partner.

Declaration of Competing Interest

We have nothing to declare.

Acknowledgements

Thank you to Medical Detection Dogs for facilitating the study and to all the participants who took part in the study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.applanim.2026.107178.

Data availability

The anonymised data collected are available as open data via the University of Bristol online data repository: https://data.bris.ac.uk/ (URLs/accession numbers/DOIs will be available after acceptance of the manuscript).

References

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beebe, S.C., Howell, T.J., Bennett, P.C., 2016. Using scent detection dogs in conservation settings: a review of scientific literature regarding their selection. Front. Vet. Sci. 3. https://doi.org/10.3389/fvets.2016.00096.

Caldicott, L., Zulch, H.E., Pike, T.W., Wilkinson, A., 2023. Olfactory Learning and Training Methods. In: Lazarowski, L. (Ed.), Olfactory Research in Dogs. Springer International Publishing, pp. 177–204. https://doi.org/10.1007/978-3-031-39370-9_9.

Christensen, R.H.B. (2022). *ordinal---Regression Models for Ordinal Data*. In (Version 2023.12-4) https://CRAN.R-project.org/package=ordinal.

Clark, C.C.A., Rooney, N.J., 2021. Does benchmarking of rating scales improve ratings of search performance given by specialist search dog handlers? Front. Vet. Sci. 8, 545398. https://doi.org/10.3389/fvets.2021.545398.

Concha, A., 2023. Detection of Human Diseases for Medical Diagnostics. In: Lazarowski, L. (Ed.), Olfactory Research in Dogs. Springer International Publishing, pp. 291–331. https://doi.org/10.1007/978-3-031-39370-9_12.

Concha, A., Mills, D., Feugier, A., Zulch, H., Guest, C., Harris, R., Pike, T., 2014. Using sniffing behavior to differentiate true negative from false negative responses in trained scent-detection dogs [Article]. Chem. Senses 39 (9), 749–754. https://doi.org/10.1093/chemse/bju045.

Cornu, J.N., Cancel-Tassin, G., Ondet, V., Girardet, C., Cussenot, O., 2011. Olfactory detection of prostate cancer by dogs sniffing urine: a step forward in early diagnosis. Eur. Urol. 59 (2), 197–201. https://doi.org/10.1016/j.eururo.2010.10.006.

Defence Science and Technology Laboratory (DSTL). (2019). Maintaining the operational performance of detection dogs. In.

Defence Science and Technology Laboratory (DSTL). (2022). Using training records to optimise detection dog performance. In.

Edwards, T.L., Browne, C.M., Schoon, A., Cox, C., Poling, A., 2017. Animal olfactory detection of human diseases: guidelines and systematic review. J. Vet. Behav. 20, 59–73. https://doi.org/10.1016/j.jveb.2017.05.002.

Edwards, T.L., Giezen, C., Browne, C.M., 2022. Influences of indication response requirement and target prevalence on dogs' performance in a scent-detection task. Appl. Anim. Behav. Sci. 253. https://doi.org/10.1016/j.applanim.2022.105657.

Essler, J.L., Wilson, C., Verta, A.C., Feuer, R., Otto, C.M., 2020. Differences in the search behavior of cancer detection dogs trained to have either a sit or stand-stare final response. Front. Vet. Sci. 7, 118. https://doi.org/10.3389/fvets.2020.00118.

Furton, K., Greb, J., Holness, H., 2010. The scientific working group on dog and orthogonal detector guidelines (SWGDOG). Natl. Crim. Justive Ref. Serv. *155*.

Furton, K., Myers, L., 2001. The scientific foundation and efficacy of the use of canines as chemical detectors for explosives. Talanta 54 (3), 487–500. https://doi.org/10.1016/s0039-9140(00)00546-4.

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

Gazit, I., Goldblatt, A., Grinstein, D., Terkel, J., 2021. Dogs can detect the individual odors in a mixture of explosives. Appl. Anim. Behav. Sci. 235. https://doi.org/10.1016/j.applanim.2020.105212.

Grandjean, D., Sarkis, R., Tourtier, J.-P., Julien-Lecocq, C., Benard, A., Roger, V., Levesque, E., Bernes-Luciani, E., Maestracci, B., Morvan, P., Gully, E., Berceau-Falancourt, D., Pesce, J.-L., Lecomte, B., Haufstater, P., Herin, G., Cabrera, J., Muzzin, Q., Gallet, C.,…Desquilbet, L. (2020). Detection dogs as a help in the detection of COVID-19 Can the dog alert on COVID-19 positive persons by sniffing axillary sweat samples ? Proof-of-concept study. https://doi.org/10.1101/2020.06.03.132134.

Guest, C., Dewhirst, S.Y., Lindsay, S.W., Allen, D.J., Aziz, S., Baerenbold, O., Bradley, J., Chabildas, U., Chen-Hussey, V., Clifford, S., Cottis, L., Dennehy, J., Foley, E., Gezan, S.A., Gibson, T., Greaves, C.K., Kleinschmidt, I., Lambert, S., Last, A., Team, C. D. R, 2022. Using trained dogs and organic semi-conducting sensors to identify asymptomatic and mild SARS-CoV-2 infections: an observational study. J. Travel. Med. 29 (3). https://doi.org/10.1093/jtm/taac043.

Guest, C., Harris, R., Sfanos, K.S., Shrestha, E., Partin, A.W., Trock, B., Mangold, L., Bader, R., Kozak, A., McLean, S., Simons, J., Soule, H., Johnson, T., Lee, W.-Y., Gao, Q., Aziz, S., Stathatou, P.M., Thaler, S., Foster, S., Mershin, A., 2021. Feasibility of integrating canine olfaction with chemical and microbial profiling of urine to detect lethal prostate cancer. PLoS. One 16 (2), e0245530. https://doi.org/10.1371/journal.pone.0245530.

Guest, C.M., Harris, R., Anjum, I., Concha, A.R., Rooney, N.J., 2020. A lesson in standardization - subtle aspects of the processing of samples can greatly affect dogs' learning (Article). Front. Vet. Sci. 7, 525. https://doi.org/10.3389/fvets.2020.00525.

Jantorno, G.M., Xavier, C.H., Melo, C.B.D., 2020. Narcotic detection dogs: an overview of high-performance animals. Ciência Rural. 50 (10). https://doi.org/10.1590/0103-8478cr20191010.

Jezierski, T., Adamkiewicz, E., Walczak, M., Sobczynska, M., Gorecka-Bruzda, A., Ensminger, J., Papet, E., 2014. Efficacy of drug detection by fully-trained police dogs varies by breed, training level, type of drug and search environment. Forensic Sci. Int. 237, 112–118. https://doi.org/10.1016/j.forsciint.2014.01.013.

Jezierski, T., Walczak, M., Ligor, T., Rudnicka, J., Buszewski, B., 2015. Study of the art: canine olfaction used for cancer detection on the basis of breath odour. Perspectives and limitations. J. Breath. Res. 9 (2), 027001. https://doi.org/10.1088/1752-7155/9/2/027001.

Kassambara, A. (2023). *ggpubr: 'ggplot2' Based Publication Ready Plots*. In (Version 0.4.0) https://cran.r-project.org/web//packages/ggpubr/ggpubr.pdf.

Krichbaum, S., Smith, J.G., Angle, C., Waggoner, P., Lazarowski, L., 2023. Sources of Human Bias in Canine Olfactory Research. Springer International Publishing, pp. 119–127. https://doi.org/10.1007/978-3-031-39370-9_6.

Kuhn, M., 2008. Building predictive models in r using the caret package. J. Stat. Softw. 28 (5), 1–26. https://doi.org/10.18637/jss.v028.i05.

Kulgod, S., Patil, B.R., Kallappa, S., Ramesh, R.S., Kulkarni, K., Somashekhar, S.P., Majumdar, S., Singh, A., Guest, C., Harris, R., Aviram, I., Shanbhag, S., Parashar, A., Ramaswamy, S.S., Bitan, I., Kulgod, A., 2026. Canine olfaction combined with Bayesian modeling for multicancer detection from breath samples: a phase ii study in India. *J. Clin. Oncol. *. https://doi.org/10.1200/JCO-25-02310.

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33 (2), 363–374. https://doi.org/10.2307/2529786.

Larmarange, J. (2023). *ggstats: Extension to 'ggplot2' for Plotting Stats*. In (Version 0.5.1) https://CRAN.R-project.org/package=ggstats.

Lazarowski, L., Dorman, D.C., 2014. Explosives detection by military working dogs: olfactory generalization from components to mixtures. Appl. Anim. Behav. Sci. 151, 84–93. https://doi.org/10.1016/j.applanim.2013.11.010.

Lazarowski, L., Krichbaum, S., Degreeff, L.E., Simon, A., Singletary, M., Angle, C., Waggoner, L.P., 2020. Methodological COnsiderations in Canine Olfactory Detection Research. Front. Vet. Sci. 7. https://doi.org/10.3389/fvets.2020.00408.

Lorenzo, N., Wan, T., Harper, R.J., Hsu, Y.-L., Chow, M., Rose, S., Furton, K.G., 2003. Laboratory and field experiments used to identify Canis lupus var. familiaris active odor signature chemicals from drugs, explosives, and humans. Anal. Bioanal. Chem. 376 (8), 1212–1224. https://doi.org/10.1007/s00216-003-2018-7.

Mancini, C., Harris, R., Aengenheister, B., Guest, C., 2015. Re-Centering Multispecies Practices: A Canine Interface for Cancer Detection Dogs Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Republic of Korea. https://doi.org/10.1145/2702123.2702562.

McCulloch, M., Jezierski, T., Broffman, M., Hubbard, A., Turner, K., Janecki, T., 2006. Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers. Integr. Cancer Ther. 5 (1), 30–39. https://doi.org/10.1177/1534735405285096.

Melchers, K.G., Lienhardt, N., Von Aarburg, M., Kleinmann, M., 2011. Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. Pers. Psychol. 64 (1), 53–87. https://doi.org/10.1111/j.1744-6570.2010.01202.x.

Noonan, L.E., Sulsky, L.M., 2001. Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a canadian military sample. Human. Perform. 14 (1), 3–26. https://doi.org/10.1207/S15327043HUP1401_02.

Oxley, J.C., Waggoner, L.P., 2009. Detection of Explosives by Dogs. Elsevier, pp. 27–40. https://doi.org/10.1016/b978-0-12-374533-0.00003-9.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. In http://www.R-project.org/.

Rooney, N., Trivedi, D.K., Sinclair, E., Walton-Doyle, C., Silverdale, M., Barran, P., Kunath, T., Morant, S., Somerville, M., Smith, J., Jones-Diette, J., Corish, J., Milne, J., Guest, C., 2025. Trained dogs can detect the odor of Parkinson's disease. J. Park. Dis. 15 (6), 1111–1115. https://doi.org/10.1177/1877718X251342485.

Sarkis, R., Lichaa, A., Mjaess, G., Saliba, M., Selman, C., Lecoq-Julien, C., Grandjean, D., Jabbour, N.M., 2021. New method of screening for COVID-19 disease using sniffer dogs and scents from axillary sweat samples. J. Public. Health. https://doi.org/10.1093/pubmed/fdab215.

Schlientz, M.D., Riley-Tillman, T.C., Briesch, A.M., Walcott, C.M., Chafouleas, S.M., 2009. The impact of training on the accuracy of Direct Behavior Ratings (DBR). Sch. Psychol. Q. 24 (2), 73–83. https://doi.org/10.1037/a0016255.

Sonoda, H., Kohnoe, S., Yamazato, T., Satoh, Y., Morizono, G., Shikata, K., Morita, M., Watanabe, A., Morita, M., Kakeji, Y., Inoue, F., Maehara, Y., 2011. Colorectal cancer screening with odour material by canine scent detection. Gut 60 (6), 814–819. https://doi.org/10.1136/gut.2010.218305.

Tami, G., Gallagher, A., 2009. Description of the behaviour of domestic dog (Canis familiaris) by experienced and inexperienced people. Appl. Anim. Behav. Sci. 120 (3-4), 159–169. https://doi.org/10.1016/j.applanim.2009.06.009.

Wasser, S.K., Davenport, B., Ramage, E.R., Hunt, K.E., Parker, M., Clarke, C., Stenhouse, G., 2004. Scat detection dogs in wildlife research and management: application to grizzly and black bears in the Yellowhead Ecosystem, Alberta, Canada. Can. J. Zool. 82 (3), 475–492.

Wickham, H., 2016. Ggplot2: Elegant graphics for data analysis, 2. Springer International Publishing. https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf.

Wilson, C., Morant, S., Kane, S., Pesterfield, C., Guest, C., Rooney, N., 2019. An owner-independent investigation of diabetes alert dog performance [Article] (Article ARTN). Front. Vet. Sci. 6, 91. https://doi.org/10.3389/fvets.2019.00091.

Wilson, C.H., Kane, S.A., Morant, S.V., Guest, C.M., Rooney, N.J., 2020. Diabetes alert dogs: Objective behaviours shown during periods of owner glucose fluctuation and stability. Appl. Anim. Behav. Sci. 223. https://doi.org/10.1016/j.applanim.2019.104915.

Zubedat, S., Aga-Mizrachi, S., Cymerblit-Sabba, A., Shwartz, J., Leon, J.F., Rozen, S., Varkovitzky, I., Eshed, Y., Grinstein, D., Avital, A., 2014. Human–animal interface: the effects of handler's stress on the performance of canines in an explosive detection task. Appl. Anim. Behav. Sci. 158, 69–75. https://doi.org/10.1016/j.applanim.2014.05.004.