Yann YVINEC[1]
Pascal DRUYTS[2]

A SIMPLE PROTOCOL FOR A DOUBLE-BLIND TEST ON AN EXPLOSIVES/DRUGS LONG-RANGE DETECTOR

1. Abstract

Many long-range explosives detectors are available on the market but their performance is subject to controversy. Very few independent test reports are public and potential purchasers have little more than hearsay and demonstrations by vendors to decide whether or not they should acquire these products. This paper proposes a simple and efficient protocol based on a double-blind test and a statistical analysis to assess whether these products are worth further investigation. A hypothesis test is defined to reject detectors that do not reach a required probability of correct claims to be chosen according to the operational requirements. The paper also presents a procedure that allows one to define the number of test samples to use in order to ensure that the hypothesis test will be able to reject a random detector. The approach is tested on a commercial long-range drugs detector. The detector is shown to be compatible with a random detector and according to the hypothesis test the detector must be rejected if the required probability of correct claims is higher than 66%, which will probably be the case for most applications.

2. Introduction

Remote detection of certain substances can be useful in many activities, such as security, police work, landmine detection, forensic, etc. Substance to detect can range from explosives, to metal, drugs, human remains, and many others. Over the decades many specific detectors have been designed for these purposes. Explosives can be detected by dogs ([1]) or by Remote Explosive Scent tracing (REST) ([8]) Dogs are also used to detect drugs. Metal can be detected by metal detectors ([7]) Ground penetrating radars ([4]) have been used to detect hidden pipes, archaeological remains, or buried human bodies.

Methods to detect a wide range of substances have been proposed but have raised scepticism among scientists: for instance dowsing, either by a pendulum or a divining rod, has been proposed for a long time.

Several companies have recently marketed long-range detectors of a vast range of substances, usually including explosives and drugs, with an operating mode similar to divining rods. These detectors raise several questions. First how do they really work? A study of documentations produced by manufacturers or vendors, and discussions with

---

[1] Royal Military Academy, Department CISS, Brussels, Belgium, yann.yvinec@rma.ac.be
[2] Royal Military Academy, Department CISS, Brussels, Belgium, pascal.druyts@rma.ac.be

some manufacturers show that manufacturers usually describes the principles of these tools in a vague or unconvincing manner, if at all. Some manufacturers claim that their products work on a same principles as divining rods or propose more scientific-looking principles. Some others may not know themselves how their products work. They may still study the question, propose some theory (one spoke about '*new physics*') or even claim their ignorance.

Explanations proposed by manufacturers include: '*electrostatic magnetic attraction*' ([5]) or '*Magneto-Electrostatic Detection*' ([12]). Sometimes the explanation is vague and limited to one phrase such as '*interaction between the detector and the substance*' or claiming that the principle is '*similar*' to magnetic resonance imagery ([6]). Some of these explanations contain scientific errors, such as confusion between the notions of potential, field, and charge. They usually fail to explain how a specific substance can be detected and not confused with others.

To the authors' knowledge no scientific study of the working principles of these tools have produced reasons to explain their performance as claimed by their manufacturers. For instance, X-rays have shown that some of these devices contain no electronics. The ideomotor effect has been suggested to explain how operators could unconsciously influence detection ([10]).

Although there seems to be no scientific explanations for the detector's claimed performance, it does not mean that it does not work. We can still determine its performance scientifically and statistically.

Few tests of these detectors have been made public ([2]). Sometimes tests have been performed by military organisation and the results are classified. Potential purchasers are left with little more than hearsay to choose to buy such tools or not.

This paper described a simple protocol to assess the detection performance of long-range detectors. This protocol can be implemented with very few resources. The objective is to help determine the performance in terms of detection in optimal conditions. This may be seen as an entry-test, or a pre-test (or pre-trial) assessment (PTA) in which a candidate detector can be examined to determine whether there is merit in committing the necessary resources to full performance tests. This may be viewed as an opportunity to exclude detectors that are clearly inappropriate. This PTA is based on a simple trial that can be repeated enough times to get statistically relevant results. This is different from a more operational test where long preparation is needed to test a detector in realistic situations, but which does not provide statistically relevant results. If this PTA is passed, other tests can be performed before purchasing the detector. Many other factors may influence the final decision: the cost, the weight, the operational procedure, the accuracy, the range, the sensitivity, the user-friendliness, etc.

3. The test protocol

The objective of the proposed test is to evaluate the ability of a long-range detector to detect a given substance in ideal conditions. It requires at least three persons: a trained operator for the detector, a sample carrier and a sample manager to prepare the samples.

One type of substance is to be used. The type and quantity of the substance is defined before the test in order to reach optimal detection capability. This choice can be

done in agreement with the manufacturer. A sample of the substance may be used to calibrate the detector if needed.

The sample is placed in a labelled plastic box.

A similar box holds an inert material. The two boxes with their samples in them should ideally be almost undistinguishable: same weight, same weight distribution, same noise when moved, etc. An easy way to prevent any noise or scent is to use a sealed, airtight box and fix the substance in it so that it cannot move.

The operator chooses an optimal position to place the samples in order to ensure an optimal detection. This is important because some detectors are said to have a detection range of several hundreds of metres depending on the amount of substance to detect. If so, unwanted signals may occur in some directions or locations.

A list of trials is prepared beforehand by the sample manager who keeps it secret. Each list item is numbered and describes what will be presented to the operator in an opaque container (target substance/inert). An efficient way to produce this list is to use a random generator to generate a uniformly distributed number between zero and one for each sample and the substance to detect if this number is larger than a chosen probability p_T. As shown in the next section, for an assessment based on the probability of correct claim, it is recommended to use $p_T = 0.5$. The number of sample to use is discussed in the next section.

A typical list would be:

Table 1 Example of list if trials

Container #	1	2	3	4	…
Substance to be detected	YES	NO	NO	YES	…
Declaration					

For each list item:
- The sample manager secretly places the appropriate plastic box (substance/inert, according to the list) in an opaque container and puts the container outside the sample room, without any contact with people on the test area. The preparation time should be similar whether the container contains the substance or not. Care should be taken to choose a container and a handling methodology to ensure that the inert containers are not contaminated by the substance to detect. This could be defined in agreement with the manufacturers. If the operator always gets an alarm because of contamination, the test should be cancelled.
- The sample carrier takes the opaque container and places it at the known, pre-defined test location.
- The operator operates the detector and, following a clearly defined procedure, declares if he or she considers that the container contains the substance or not. He or she may move to several positions but must provide an answer (yes or no) within a reasonable time decided in advance. The authors have participated to a test of such a detector where the operator could not still decide after 40 minutes if there was a

substance or not. This is clearly unacceptable in the proposed test and for most operational contexts.

- − The observer communicates the operator's claim (substance/inert) to the sample manager but restricts further communications to the strict minimum to avoid exchange of information between test site and sample room.
- − The sample manager writes down the operator's claim.
- − The sample carrier brings the container to the door of the sample room and goes back to the test site
- − The sample manager takes the container inside the sample room, removes the opaque container and checks the box label (substance/inert) to crosscheck with the list.

The procedure is repeated with the next sample, according to the list.

Once all containers have been tested the results and the contents of the containers are compared. A copy of the list is provided to the participants.

Before the evaluation the operator does not know how many containers will hold drugs.

Additional observers, including from the manufacturers, may be present but they may not circulate between the test-site to the sample room, as contacts between these two locations must be avoided (even during breaks). Measures should be taken to reduce preparation or recording errors.

4. Statistical analysis

4.1. Designing the test

The test results may be summarised as follows:

Table 2 Example of presentation of results (*a*, *b*, *c* and *d* are the number of claims in each category)

		Actual	
		Substance to detect	Inert
Claims	**Substance to detect**	a	b
	Inert	c	d

The ratio of correct claim is given by $(a+d)/(a+b+c+d)$. The ratio of correct detection is given by $a/(a+c)$. The false alarm rate is $b/(b+d)$. The ratio of number of samples with the substance to detect to the total number of samples is $f_T = (a+c)/(a+b+c+d)$.

Choosing f_T may have an effect on the probability of correct claims. To assess this, let the detector under test have a probability p_D of detecting a sample containing the substance to detect (detection probability) and a probability p_{FA} of having a signal when the sample does not contain the substance to detect (false alarm rate). Let's assume that these two values are constant. Ideally the detection probability should be close to one and the false alarm rate should be close to zero.

The probability to have a correct claim is therefore:

$$p_{CC} = f_T p_D + (1 - f_T)(1 - p_{FA})$$

(1)

For an ideal detector, this probability is equal to 1.

Let's consider a random detector that produces a random alarm with a probability p_0 regardless of whether the container contains the substance to or not. Then its detection probability and its false alarm rate are both equal to p_0.

Introducing this in (1) yields:

$$p = p_0(2p_T - 1) + 1 - p_T$$

(2)

For this probability to be independent of p_0 the substance to detect should be presented to the operator with a ratio of 0.5 and the probability of correct claim then equals 0.5, which is then a reference against which to compare the ratio of correct claim for any detector. We therefore recommend using $f_T = 0.5$. Note that the method proposed in the previous section easily allows fixing p_T. This is, however, expected to have little influence, especially for large a number of samples.

4.2. Confidence intervals

The ratios defined above are samples of random variables. There are only estimations of the underlying probabilities. In order to better assess these probabilities, confidence intervals should be used.

With the assumption on the detection presented in 4.1, consider a number of correct claims of n_{CC} out of n trials. We want a confidence interval for the probability of correct claim characterised by a value α as follows.. We choose as lower bound of the interval the probability of correct claim of detector such that the probability to obtain a number of correct claims lower than n_{CC} is (1- α)/2 and the upper bound of the interval the probability of correct claim of a detector such that the probability to obtain a number of correct claims larger than n_{CC} is (1- α)/2. The interval is then the Clopper-Pearson 100(1- α)%-confidence interval.

The lower and upper bounds of the confidence interval for the probability of detection can be computed as follows ([11], [3]).

$$P_L = \frac{v_1 F(v_1, v_2, \alpha/2)}{v_2 + v_1 F(v_1, v_2, \alpha/2)} \quad \text{with } v_1 = 2n_{CC}, v_2 = 2(n - n_{CC} + 1)$$

$$P_U = \frac{v_3 F(v_3, v_4, 1 - \alpha/2)}{v_4 + v_3 F(v_3, v_4, 1 - \alpha/2)} \quad \text{with } v_3 = 2(n_{CC} + 1), v_4 = 2(n - n_{CC})$$

(3)

where P_L is the lower bound of the confidence interval for the probability of detection, P_U is the upper bound of the confidence interval for the probability of detection, and $F(f, g, \lambda)$ is the λ-quantile of the F-distribution with f and g degrees of freedom.

The above equations are undetermined for n_{CC} equal to zero or n. If there is no correct claims ($n_{CC} = 0$) then the following bounds are:

$$P_L = 0 \qquad\qquad\qquad\qquad\qquad (4)$$
$$P_u = 1 - \sqrt[n]{\alpha}$$

If all claims are correct ($n_{CC} = n$) then the following bounds are:

$$P_L = \sqrt[n]{\alpha} \qquad\qquad\qquad\qquad\qquad (5)$$
$$P_U = 1$$

For the confidence interval of the probability of detection, the same formula hold by changing *n* by *a+c* and n_{CC} by *a*. For the confidence interval of the false alarm rate, change n by *b+d* and n_{CC} by *b*.

A first method to analyse the results is therefore to see if the frequency 0.5 belongs to the 95%-confidence intervals. If it does, it means that the assumption of a random detector is compatible with the data. This does not mean, however, that the detector works randomly. For instance, if few trials are made, the 95%-confidence interval will be very large and may contain the frequency 0.5, how good the detector may be. It is therefore important to make enough trials to reduce the size of the confidence interval. This can be further quantified by use of a test hypothesis as described in the next section.

4.3. Hypothesis test

With a statistic test, it is impossible to prove that a detector performs as a random detector. As explained above it is possible to say that its performances are compatible with those of a random detector. In practice, however, a detector is only appropriate if it has a certain level of performance that should be determined according to the intended usage. The performance may be quantified with a number of indicators such as the probability of detection, the probability of false alarms or the probability of correct claims. In this paper we will only consider the probability of correct claims and we will define a hypothesis test that can be used to reject a detector if the detector has not the required probability of correct claims. Hypothesis test based on probability of detection and probability of false alarm may also be useful for a number of applications. This will be the subject of further research. Note, however, that it makes no sense to consider probability of detection or probability of false alarm alone. The requirement should always be on a combination of both. Indeed, claiming detection for each sample leads to 100% detection.

The proposed hypothesis test is based on two hypotheses. H_0: "The detector has a probability of correct claim larger or equal to $p_{CC}^{\min}$" and H_1: "The detector has a probability of correct claim smaller than $p_{CC}^{\min}$". The two hypotheses are contradictory and the detector should be rejected if hypothesis H_1 is retained.

Here again, it is impossible to prove that H_1 or H_0 is true but the objective of the hypothesis test is to reject the detector only if H_0 is unlikely to be true. In other words, doubt benefits to the detector and the test is design to ensure that an inappropriate detector is rejected.

This is quantified by the power of the test $1-\beta$ with β the probability that an appropriate detector is rejected by the test (false negative rate).

The test is then performed by rejecting H_0 (rejecting the detector) if $n_{CC} < n_{CC}^{\min}$ where n_{CC} is the number of correct claims and $n_{CC}^{\min}$ is a threshold computed to ensure that an appropriate detector would be rejected with only a small probability β.

We first consider the probability that the detector has a probability of correct claims p_{CC} if there are n_{CC} correct claims.

$$p\left(p_{CC}|n_{CC}\right) = \frac{p\left(n_{CC}|p_{CC}\right)p\left(p_{CC}\right)}{p\left(n_{CC}\right)} \tag{6}$$

We consider a uniform distribution for p_{CC}. The probability that a detector with a probability of correct claim p_{CC} gives n_{cc} correct answers in n trials is given by the binomial model:

$$p\left(n_{CC}|p_{CC}\right) = C_{n_{cc}}^{n} p_{CC}{}^{n_{cc}}\left(1 - p_{CC}\right)^{n - n_{cc}} \tag{7}$$

The probability that a detector that provides n_{CC} correct claims out of n has a probability of being correct smaller than $p_{CC}^{\min}$ is:

$$p\left(p_{CC} \le p_{CC}^{\min}|n_{CC}\right) = \frac{\displaystyle\int_{0}^{p_{CC}^{\min}} p_{CC}{}^{n_{cc}}\left(1 - p_{CC}\right)^{n - n_{cc}} d\alpha}{\displaystyle\int_{0}^{1} p_{CC}{}^{n_{cc}}\left(1 - p_{CC}\right)^{n - n_{cc}} dp_{CC}} = B\left(p_{CC}^{\min}; n_{CC} + 1, n - n_{CC} + 1\right) \tag{8}$$

where B is the incomplete beta function ([9]).

Then the detector can be rejected with a confidence $1-\beta$ if $B\left(p_{CC}^{\min}; n_{CC} + 1, n - n_{CC} + 1\right) > 1 - \beta$. A value of 0.05 can be chosen for β. If a detector is not rejected, it does not necessarily means that there is a large confidence in its probability of correct claim being larger than $p_{CC}^{\min}$, because this depends on the number of trials in the test.

4.4. Determination of the number of trials

In this section we discuss how the number of test samples should be chosen if we suspect a random detector and we want to have good chances to reject it with the hypothesis test presented in the previous section. That is, we want to reject the detector (with a given confidence, say 95%) if its probability of correct claim is lower than a threshold $p_{CC}^{\min}$, which is chosen according to operational considerations. Obviously the higher $p_{CC}^{\min}$ is, the less measurements are needed to reject a random detector. Equation (8) can be used to compute the number of trials required to be just able to reject a random detector with a given $p_{CC}^{\min}$ and a given confidence but this requires the ratio of correct claims to be known.

This ratio is not known before the test, it is a random variable. The number of correct claims follows a binomial distribution. The most likely value is

$$n_{CC}^{ML} = \frac{n}{2} \tag{9}$$

Using this value, the limit $p_{CC}^{\min}$ can be computed as a function of n (Fig. 1, curve with crosses). Using n_{CC}^{ML} may be dangerous, because the number of correct claims obtained for a given test will be higher than n_{CC}^{ML} with a probability of 0.5. If $n_{CC} > n_{CC}^{ML}$ and n is chosen according to Equation (8) with n_{CC}^{ML} the hypothesis test will not reject a random detector. A more conservative approach is to compute n using $n_{CC,95}$ defined by (Fig. 1, curve with circles):

$$p(n_{CC} < n_{CC,95}) = 0.95 \tag{10}$$

With such a choice, there is a probability of 0.95 that the number of correct claims obtained is smaller than $n_{CC,95}$ and, for all these cases, a random detector will be rejected with the chosen number of tests. As an example, on sees (Fig. 1) that if $p_{CC}^{\min} = 0.7$, the number of tests required if n_{CC}^{ML} is used is 20 and the more conservative number of tests obtained using $n_{CC,95}$ is 60.
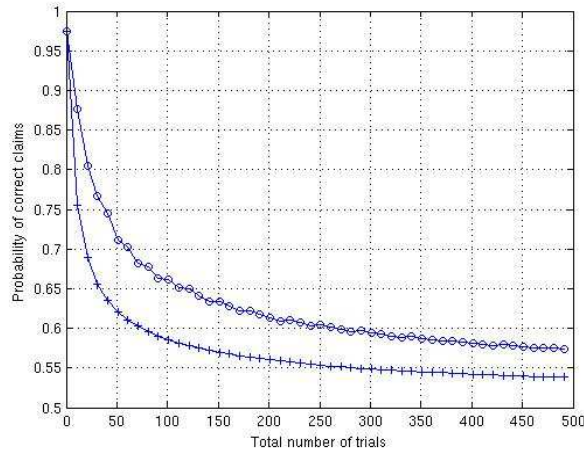
Fig. 1 Thresholds of probability of correct claim that can be used to reject a detector with a 95% confidence, as a function of the number of trials; See text for details (the curve is not smooth due to integer computation)

5. Example

We have used the proposed protocol to test a commercial long-range detector. This detector is equipped with cards designed to detect specific substances. The detector works with special cards for the detection of specific substance. Since only cards for drugs was available, the test involved only drugs samples.

Sixty-six samples were presented to an operator. The samples were either a box of seven grams to ten grams of cannabis (26 times out of 66) or a box of small pieces of

paper (40 times out of 66). For each sample the operator gave a clear and unambiguous claim.

The results are summarised in the following table.

Table 3 Results of a double-blind test of ADE 651

		Reality		
		Drugs	**No drugs**	**Total**
Claims	**Drugs**	10	14	**24**
	No drugs	16	26	**42**
	Total	**26**	**40**	**66**

We can see that 55% of the claims were correct (36 out 66), 38% of the samples containing drugs were detected (10 out of 24) and 35% of the samples that did not contained drugs were mislabelled as drugs (14 out of 40).

The 95%-confidence interval of the probability of correct claim is (42%–67%), for the detection probability is (20%–59%) and for the false alarm rate (21%–52%).

It appears that the assumption of a random detector is compatible with the data.

Suppose now that we want to reject the detector if its probability of correct claim is lower than, say, 0.7. Then by using, $p_{CC}^{\min} = 0.7$, $n = 66$ and $n_{CC} = 36$ in Equation (8), we obtain

$$p\left(p_{CC} < p_{CC}^{\min}\middle|n_{CC}\right) = 0.9964 = 1 - 0.0036 \tag{11}$$

This indicates that the probability to make an error when rejecting this detector is 0.0036 (or 0.36%).

On the other hand, with $\beta = 0.05$, the detector must be rejected if the lowest acceptable probability of correct claim is

$$p_{CC}^{\min} = B^{-1}\left(1 - \beta; n_{CC} + 1, n - n_{CC} + 1\right) = 0.64 \tag{12}$$

where B^{-1} is the inverse incomplete beta function.

6. Conclusions

This test protocol presented here focuses on performance on the detector in optimal conditions and serves only to make a first rejection. A hypothesis test has been defined to reject detectors that do not reach a required probability of correct claims, a probability to be chosen according to the operational requirements. We also presented a procedure that allows one to define the number of test samples to use in order to ensure that the hypothesis test will be able to reject a random detector. The approach is tested on a commercial long-range drugs detector. The detector is shown to be compatible with a random detector and according to the hypothesis test the detector must be rejected if the required probability of correct claims is higher than 66%, which will probably be the case for most applications.

References

[1] BACH, MCLEAN, AAKERBLOM, SARGISSON Improving mine detection dogs: an overview of the GICHD dog program. Article by published in the Proceedings of EUDEM2-SCOT - 2003, volume 1, 149 - 155 (September 2003)

[2] BANKS G. Test report: The Detection Capability of the SNIFFEX handheld Explosives Detector, Naval EOD technology Division, US, 2005

[3] CEN, CEN Workshop Agreement (CWA) 14747-2:2008: Humanitarian Mine Action – Test and Evaluation – Soil characterisation for metal detector and ground penetrating radar performance, European Committee for Standardization, 2008

[4] DANIELS D. Ground penetrating radar, 2nd edition, Edited by David Daniels, 2004

[5] DEVLIN H. ADE651 bomb detector? Convincing jargon — unconvincing product, The Times, 28 November, 2009

[6] EMCOM AFRICA, Comstract – Alpha 6 Molecular Detector, accessed from www.emcom.co.za/products:comstract.html, accessed on April 2010.

[7] GUELLE D. SMITH A., LEWIS A.BLOODWORTH T. The metal detector handbook, European Communities, 2003

[8] MCLEAN I.G. and SARGISSON R. Optimising the Use of REST for Mine Detection, The Journal of Mine Action 8.2, November 2004

[9] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, April 2010

[10] RANDI J. The Matter of Dowsing, Swift, vol 2, No 3&4, James Randi Educational Foundation, 1998

[11] SIMONSON K M, Statistical Considerations in Designing Tests of Mine Detection ystems: I — Measures Related to the Probability of Detection, technical report, andia National Laboratories, August 1998

[12] UNIVAL-GROUP Presentation: Handheld Explosive Detection device HEDD1, accessed on www.hedd1.com on April 2010.

A SIMPLE PROTOCOL FOR A DOUBLE-BLIND TEST ON AN EXPLOSIVES/DRUGS LONG-RANGE DETECTOR

Summary in English

Many long-range explosives detectors are available on the market but their performance is subject to controversy. Very few independent test reports are public and potential purchasers have little more than hearsay and demonstrations by vendors to decide whether or not they should acquire these products. This paper proposes a simple and efficient protocol based on a double-blind test and a statistical analysis to assess how good these products are at detecting explosives, drugs or other substances. This development is based on lessons learned of actual tests performed on such detectors

This protocol was used to test a commercial long-range detector. The detector was shown to be compatible with a random detector and the hypothesis test showed that the detector could be rejected (with a 95% confidence) if the requirement was to reject detectors whose probability of correct claims were lower than 0.64.