Testing and Evaluating Results of Research in Mine Action

Yann Yvinec

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65780

Abstract

This chapter summarizes the experience of the Royal Military Academy in testing and evaluating new tools for mine action. It first underscores the importance of testing and evaluating new methods in general and in mine action in particular. Some recommendations are given to help the design of test protocols: defining carefully the objectives of the test and what is to be measured, the importance of blind and double-blind tests, choosing between realism and statistical relevance, the importance of how to display the results, etc. These recommendations are illustrated by real-life examples, mainly from test and evaluation of detectors of mines in which RMA has been involved. A test protocol is detailed. It is the one that RMA designed and used to evaluate a detector that was proven to be useless and that led to the criminal conviction of its designer in the United Kingdom. Sources of available test protocols and test reports are also listed.

Keywords: test and evaluation, test protocol, standardisation, test design

1. Problem statement

Evaluating a solution for a problem against other, state-of-the-art solutions to the same problem is an integrant part of research. In mine action, however, this is even more important because the mine action community is both anxious to use new methods to help their work and reluctant to trust new tools. Some activities of mine action, for instance the detection of mines or other explosives, may be dangerous and changing a well-know detector for a new and possibly unknown detector is not something all mine action practitioners will do very easily. One way to promote the adoption of new tools or methods is to evaluate them in a way that would convince potential users of the added value of the proposed solution.

A trial is usually an activity you perform in order to answer a given question; does my method achieve the expected results? Is it better than other, existing methods? What are its limitations? The trial must be designed to provide answers to the question.

Therefore, designing correctly a test protocol is of paramount importance.

2. Test protocols and why bothering about them

A test protocol describes how the trial must be performed. It serves several purposes.

When correctly designed it ensures that the objectives of the test will be achieved. This is essential in order not to waste the time, money and resources required for the trial.

During the trial itself the protocol indicates clearly the steps to carry out. When the trial is carried out to decide which of two options is better and some participants of the tests differ on their favourite options, it is important that they all agree on the test protocol before the test begins in order to prevent arguments after the results are known. It is usually considered as a good practice to select the criteria to decide the outcome of a test (success or failure for instance) before the start of a trial.

When included into a test report, the protocol allows readers to assess the conclusions of the report. Conclusions of test reports can be debated just because there are some concerns about the protocols that have been used. How to analyse the results is an important part of a test protocol. The authors of a study published in 1990 considered that they had empirically proven the reality of dowsing, the ability to discover underground water with instruments such a wooden stick or a pendulum. A different statistically analysis of the same data showed that, on the contrary, no such effect could be proven [1]. The whole controversy focused on the way to analyse the data.

The current controversy over the lack of reproducibility of research published in very different scientific domains is intense [2–5]. The explanations often given to questions some published results are the use of too small samples in the experimentation and a misunderstanding of statistics in the analysis of the results.

3. Some key points to keep in mind when designing a test protocol

3.1. Introduction

A good trial, that is, a trial that provides trustworthy and meaningful information, requires sufficient allotted time, sufficient funding, adequate personnel and resources. Before starting a thorough test, it could be useful to have a first, rough evaluation of the equipment to test in order to reject right away equipment that is obviously below the expected performance. This first evaluation could be a basic check when the equipment is based on new or uncertain scientific principles. Many new ideas have been proposed to detect landmines: radar, acoustic systems, thermal systems, rats, bees, etc. Some long-range detectors of explosives have been proposed and proved to be fakes. One such system was sold to many clients, although the working principles were clearly scientifically unsound. This led eventually to the conviction of the author to a prison-sentenced of 10 years [6].

Besides, a scientific evaluation of the principle, some basic tests could be performed in order to check if the equipment works in ideal conditions. Once such an entry-test is successful a full evaluation of the equipment with varying and controlled, conditions could be carried out.

A last test could be an evaluation of the equipment in operational conditions.

When devising a test protocol there are a few issues to consider.

3.2. What is the test for?

In 2007 the mine action community was surprised by the results of a test report of a comparison of the performance of metal detectors [7]. The test showed that most metal detectors had a probability of detection significantly lower than 100%. This result conflicted with the daily experience of regular users of metal detectors. The reason of these apparent strange results was due to the very specific objectives of the test as defined by the authors. The goal of the tests was not to evaluate the field performance of metal detectors but to compare them. Since most metal detectors were expected to have good performance in normal field conditions, the test was designed with extreme conditions in terms, for example, of depths at which the target mines were buried, so that the detectors could be sometimes at the limit of their detection capability. In these extreme conditions metal detectors would have performance below 100% and it was expected that the decrease would be different from one detector to the other, thus allowing a better comparison. If the metal detectors performed so poorly in these tests, it was because of the extreme conditions of the tests, not necessarily because of intrinsic poor performance. The purpose of a test had a direct impact on the test protocol, therefore on the results and hence on the way the reader should interpret the results. Unfortunately, identifying the purpose of a test is something that is sometimes overlooked by readers and test designers alike. A test designed for a given purpose may be irrelevant for another. The answer to a question is not necessarily useful for another question.

3.3. What does ‘detection’ means?

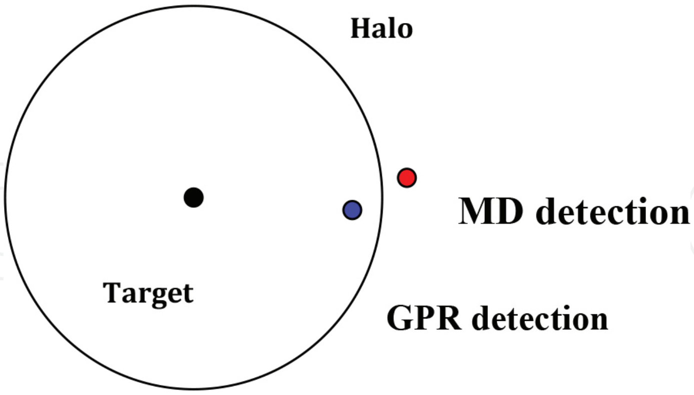
This seems like a stupid question, but when evaluating the performance of a detector, defining precisely when the operator using the detector has actually detected something is important. If an operator looks for anti-personnel mines with a metal detector and finds an alarm two meters away from the only mine present, the alarm is clearly a false alarm. But how close to the mine should the alarm indication be in order to consider that this is a true detection and nor a false alarm?

A second reason to be careful is when the operator has to make a decision based on a signal, whether it is audio, visual or otherwise. It is important to make the difference between the performance of the detector and the ability of the operator to locate precisely a target. In addition to defining properly what detection is, defining what an alarm is—whether it turns out to be true detection or a false alarm—must also be done carefully.

The guidelines to evaluate the performance of metal detectors described in T&E Protocol 14747-1 address the two problems [8]. For instance an operator should sweep the detector over the target five times *"to determine whether detection is consistently indicated"*. They also propose a method to decide if an alarm should be considered as detection or a false alarm. Detection is defined when the distance between the indication location and the actual target is shorter than a given distance value. The detection halo of a target is defined as the circle whose radius is this given distance. According to Ref. [8], B6, page 59 *"The radius of the detection halo shall be half of the maximum horizontal extent of the metal components in the target plus 100 mm"*.

This definition was designed for metal detector. It may have to be adapted for new detectors. Experience shows that other definitions of halos have been used. For instance in a report from DGA, 'precise detection' is defined as a detection within 20 cm of the target and 'close detection' when detection is within 40 cm.

In 2009 the Royal Military Academy was asked by the International Test and Evaluation Programme for Humanitarian Demining (ITEP) to send a scientist to invigilate a trial of metal detectors and dual sensors that combined metal detectors and ground-penetrating radars. Public results can be found in Ref. [9]. But an unpublished early result proved to be quite interesting. The protocol of use of the dual sensor was as follows: the operator would first use the metal detector to determine a suspicious location, mark it with a plastic chip of a certain colour and then use the ground-penetrating radar to confirm detection. In case of confirmation the operator will place a plastic of a different colour at the location. Sometimes as illustrated in **Scheme 1** results showed that the alarm indications of the metal detector were outside the defined halo but the alarm indications of the ground-penetrating radar were within.



Scheme 1. Example of a test result where the alarm indication of a metal detector (in red) is outside the halo of the target (in black) and the alarm indication of the ground-penetrating radar (in blue) used as a confirmation detector is inside.

Since the footprint of the ground-penetrating radar is smaller than the metal detector footprint—making it easier to pinpoint with the radar than with the metal detector—and some targets have metal parts at their borders and not at their centres, it was indeed possible that an operator using the dual sensor puts some metal detector chips just outside some halos and the corresponding radar chips just inside the same halos. By a strict definition of a hit the conclusion may be that the probability of detection of the radar, although used as a confirmation sensor, may be higher than the probability of detection of the metal detector.

This is counterintuitive. How to solve this problem?

One possibility would be to consider that there is a dual sensor hit if there is both a metal detector hit and a radar hit. This solution may lose the advantage of the radar, including its pinpointing ability.

Another possibility would be to consider that, if there is a radar hit, we should consider that there is a metal detector hit, even if the metal detector chip is a little outside the halo. Changing the definition on a metal detector hit in this manner may make it difficult to compare the dual sensor metal detector and the corresponding stand-alone metal detector.

Defining clearly a dual sensor hit is important. During the tests, some operators even wondered whether they should put the chips where they had a signal or where they thought the target was.

As a side note, during this test it appeared that one operator of the dual sensors was found out that he was colour-blind!

We could also consider that if the metal detector chip is just outside the halo and the radar chip is just inside, then the problem is that the halo is too small for the metal detector. The metal detector chip may have been outside the halo ('miss') but it was close enough to the target to allow the operator to find the target with the radar. It can be argued that the target was actually detected with the metal detector but the pinpointing was not accurate enough. Should the halo be increased then? If so, how to define the new halo?

If we want to increase the halo, then we may want to increase it also for metal detector trials. This may lead to a modification of Ref. [8].

In the report to the ITEP Executive Committee in Oberjettengberg, Germany, on 14 November 2009, a graph was presented showing what happens if the halo is doubled. With the original halo the ratio of detection was between 0.5 (for laterite soils) and 0.9 (with humus). For most detectors during the tests, multiplying the halo size by two led to a ratio of detection above 0.9.

It means that 90% of the targets had chips around them. Does that mean that the detection was correct but the pinpointing was inaccurate, or that there were so many chips on the lanes that when the halos are sufficiently increased most halos end up containing chips?

Increasing halos have an impact on both the probability of detection and the false alarm rate. When the halos are increased, more chips tend to be inside halos and therefore the probability of detection tends to increase. The false alarm rate is defined as the ratio of the number of false alarms to the area of the lane outside any halo. When the halos increase, both numbers tend to decrease. It is therefore difficult to predict how the ratio will change. If the halos increase too much they may intersect with each other. This may create problems for both the probability of detection and the false alarm rate. How should a chip inside two halos be considered? As one hit? And if so, for which target? Or as two hits? Intersecting halos have impact on the false alarm rate too. When halos intersect computing the area outside halos is a more difficult than if halos do not intersect. It is therefore easier to have halos that do not intersect with each other. How far can the halos be increased without having them intersecting? It depends on the smallest distance between targets.

When designing a new test site, using larger halo sizes may require increasing the smallest distance between targets.

Theoretically, the ratio of detection depends on many factors: the detectors, the targets, the soil, the operators, etc. But there is no reason why it should also depends on the halo size, which is only a tool to analyse the data. Halo sizes, however, have an impact on the number of what is defined as hits. A clear difference should therefore be made between probability of detection and the ratio of hits to the number of targets, which we can call the hit ratio or the ration of detection.

The probability of detection is what we want to estimate by the trials. Its value is unknown. During the trials we measure the hit ratio as exactly as possible. From this hit ratio and other information such as the number of targets we can estimate the probability of detection and some confidence intervals. Hit ratio is known exactly. There is no confidence interval for it. Probability of detection is estimated. So is its confidence interval.

This difference is important because the halo size, as a tool to analyse data, has no influence on the real performance of a detector in a given situation, but it has one on the hit ratio that is measured.

Figure 1 shows the hit ratio as a function of the false alarm rate for a halo size ranging from the definition in Ref. [8] to the double for three detectors that were tested. Note the scale used for the hit ratio.

When halo sizes increase, the hit ratio increases, as expected.

Sometimes the increase of hit ratio is large; sometimes it is small. There are two reasons, at least, why it could be small.

One reason is when the hit ratio is already high. With few missed targets that could become detected when halos increase, the hit ratio can only increase a little.

The other reason is when the false alarm rate is small. With few false alarms that could become hits when halos increase, the hit ratio increases only a little.

On the other hand, the case where the hit ratio could benefit the most from an increase of halo sizes is when the hit ratio is low to begin with and the false alarm rate is high.

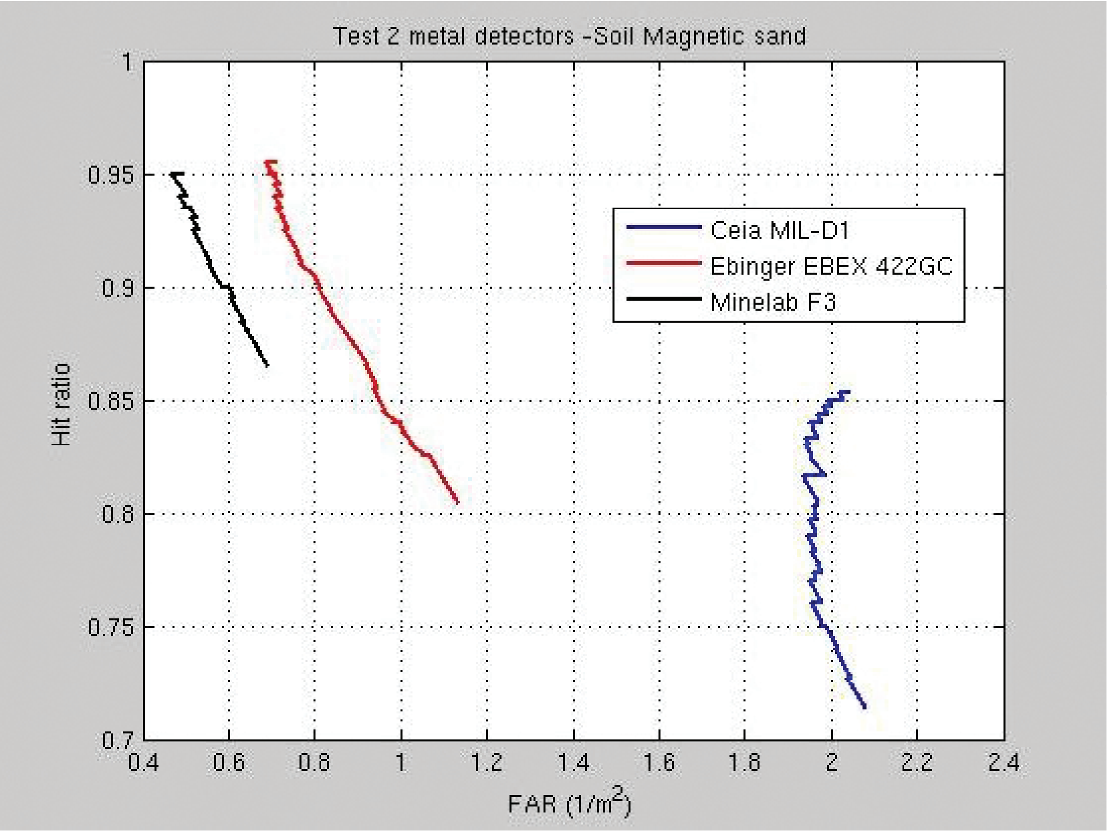


Figure 1. ROC curves as functions of the halo size.

3.4. What does 'false alarm' means?

This question is of course strongly related to the definition of detection.

When testing a detector it is not only important to estimate its capability to detect relevant target, namely its sensitivity, but also whether what it detects is a legitimate target, that is, its specificity. The false alarm rate is often used to quantify this. It can be expressed in number of false alarms per square metre, although the ratio of number of false alarms to the number of detection is sometimes used. The question then rises about the exact nature of a false alarm.

When testing a metal detector, should the presence of a small amount of metal instead of a mine be considered as a source of false alarms? For operators during clearance operations, it is a false alarm because their objective is to find mines, not scraps of metal. On the other hand, manufacturer would consider that metal detector should detect metal and then consider the detection of small metal pieces as a positive detection.

In order to solve these difficulties, [8] 8.1.2, recommends that "*the presence of metal objects around, above and under the test areas shall be avoided as far as possible*".

With other technologies a work around is not so easy. Ground-penetrating radars, for example, can detect soil disturbances. Such disturbances can happen when a test target is buried in the test areas. Therefore, in case of detection it is unclear whether the target itself or the soil disturbance triggered the alarm. But creating a test area where there is nothing *"around, above and under the test areas"* that would create a radar alarm might be extremely difficult and may lead to a test site that is not representative to reality.

If a test area contains a small number of locations that create unwanted radar alarms, these locations may be mapped. Then test targets should not be buried close to these locations. And after the trial, radar alarms close to these locations should be removed from the analysis.

Another method would be not to use local soil but soil with known characteristics, such as sand-containing grains of a given diameter. But the radar waves that would propagate through this soil would reflect at least partially at the interface between the soil and the outside of the test areas, creating false signal that might affect detection. This method might require large and deep test areas.

3.5. Should the test be blind? Double blind?

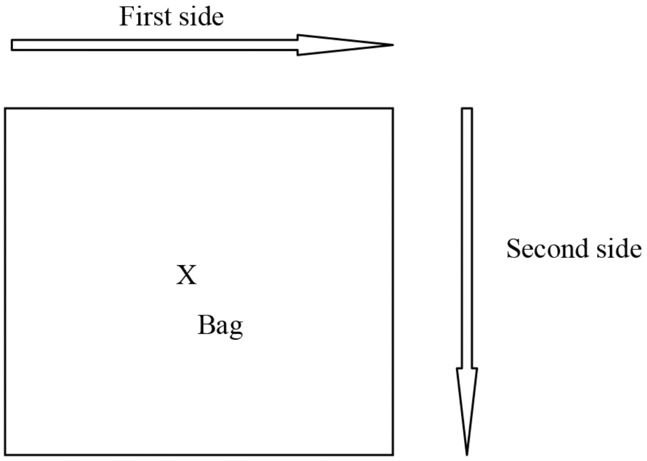
We all know that we tend to see what we want to see. For this reason, when testing a detector, the operator must not know where the targets are hidden. A test where the operator does not know the right answer to the test is said to be a blind test. Moreover, a clear separation should be made between people who analyse the results and anyone involved in the data collection of a test. Observers who where the targets are hidden may give unintentional clues to the operator about whether the operator is performing well or not. In medical study doctors checking the health of a volunteer for a test should not know whether the volunteer has received the medicine under test or a placebo. When a test is designed in such a way that anyone involved in the data collection have no knowledge that may influence their behaviour the test is said to be double blind. Double-blind tests are considered much better than blind tests.

3.6. How to deal with human factors?

Carrying out double-blind trials is a way to avoid some human factors. Having several operators for the same detector is also a possibility, although it increases the number of tests.

Human factors are all the more important when the detectors require some interpretation from operators. It could be because the sensors output an audio signal or provide complex display, such as images.

An extreme case is ADE 651, a sensor that has been proved to be totally ineffective. It is composed of a horizontal antenna that can rotate freely around a handle. When used properly it was said to point in the direction of a source of explosive, narcotics or any other sub-stance the system was calibrated to detect. During a trial supervised by the Royal Military Academy, a bag was located in a test area. The operator did not know whether the bag contained the substance to detect or something else. The operator would move around the bag along a square-like trajectory. He would first walk along the first side of the square, along a few metres. Either the antenna would point to the bag or not. Then he would move along the second side of the square, as seen in **Scheme 2**. The antenna would then either confirm the first result or not. In case of confirmation, either positive or negative, the operator would usually stop and make his claim. Then we would test a second bag. If during the second side of square the antenna would give a different indication than with the first one, the operator would try again with the third side; he would also try again the second or first side or both. After some time, he would make his claim. Almost every time his final claim was consistent with the indication of the antenna during the first side of the square. Was it coincidence or did the operator made up his mind after the first side, even if unconsciously and then try, again unintentionally, to confirm his first impression?



Scheme 2. Design of the test area for ADE 651.

3.7. Can a test be both realistic and statistically relevant?

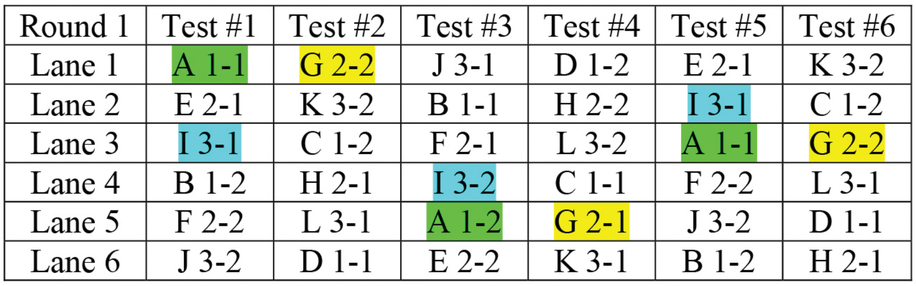
There are many parameters that have an influence on the evaluation of the performance of equipment. For a detector they include, but are not limited to, the types of targets used during the test, the environmental conditions and the operator. If you want to study the influence of these factors you may want to make tests where they vary. You may want to use different kinds of targets, locate them in different situations, use several operators, etc. When factors may vary the test designed is said to be factorial and it leads to a number of runs that can become very large. In order to get statistically relevant results, the same tests might have to be carried out several times in similar conditions in order to reduce the impact of unwanted factors. But then such a test may require a long time and huge resources, and ultimately cost a lot of money. For these reasons a statistically relevant test may be carried out on a smaller set of conditions. But then the operational relevance is questionable. On the other hand, a small scale test with few redundant data collection may be fast, cheap and give clear results but it might be difficult to guarantee the relevance of the results. A trade-off must therefore be found between realism and statistical relevance.

Several ways to tackle this problem exists. One is to clearly distinguish between factors that have a random effect, and for which several replicable tests may be required and conditions that are more deterministic and for which only a few experiments are enough. Another way is not to use all runs of a factorial design, that is, not use all combinations of all factors, but only a fraction. This is called fractional factorial design. Several options exist to choose that runs to use.

The fractional factorial design is based on a delicately balanced data set. What happens if some data are not collected because of some external reason, such as lack of time or bad weather?

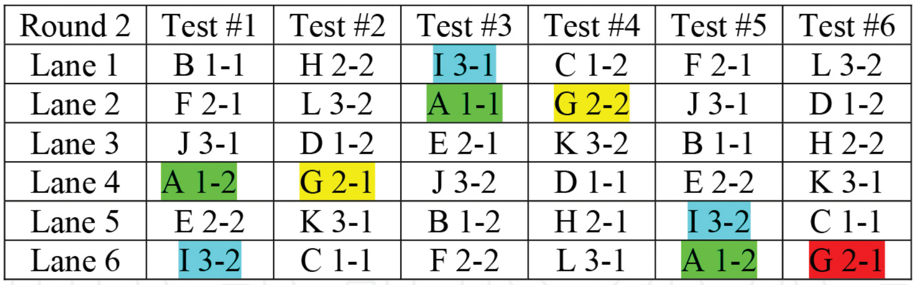
An example of fractional factorial design is given in Ref. [10] and is reproduced below.

Detectors are identified by two numbers: one is the model and the second is the specimen. For instance detector 2–3 is model 2, specimen 3. Operators are denoted by letters. Operators A, B, C and D are trained to operate detector 1. Operators E, F, G and H are trained for detector 2 and operators I, J, K and L for detector 3. **Tables 1** and **2** show the two rounds of six consecutive 'starts' in six different lanes. The document states that: "*After both rounds are executed, each detector will have been operated in each lane by all four persons trained for that detector*".



Round 1	Test #1	Test #2	Test #3	Test #4	Test #5	Test #6
Lane 1	A 1-1	G 2-2	J 3-1	D 1-2	E 2-1	K 3-2
Lane 2	E 2-1	K 3-2	B 1-1	H 2-2	I 3-1	C 1-2
Lane 3	I 3-1	C 1-2	F 2-1	L 3-2	A 1-1	G 2-2
Lane 4	B 1-2	H 2-1	I 3-2	C 1-1	F 2-2	L 3-1
Lane 5	F 2-2	L 3-1	A 1-2	G 2-1	J 3-2	D 1-1
Lane 6	J 3-2	D 1-1	E 2-2	K 3-1	B 1-2	H 2-1

Table 1. Example of a fractional factorial design (round 1).



Round 2	Test #1	Test #2	Test #3	Test #4	Test #5	Test #6
Lane 1	B 1-1	H 2-2	I 3-1	C 1-2	F 2-1	L 3-2
Lane 2	F 2-1	L 3-2	A 1-1	G 2-2	J 3-1	D 1-2
Lane 3	J 3-1	D 1-2	E 2-1	K 3-2	B 1-1	H 2-2
Lane 4	A 1-2	G 2-1	J 3-2	D 1-1	E 2-2	K 3-1
Lane 5	E 2-2	K 3-1	B 1-2	H 2-1	I 3-2	C 1-1
Lane 6	I 3-2	C 1-1	F 2-2	L 3-1	A 1-2	G 2-1

Table 2. Example of a fractional factorial design (round 2).

What would happen if one data set was missing?

Suppose Round 2, Test #6, lane 6 is missing (operator G with detector 2–1, in red in the table). On lane 6, detector 2 is no longer used by all four operators trained for it.

In order to have detector 2 used by the same number of operators on all lanes, one possibility is to remove all data sets with operator G (in yellow in the table). Then detector 2 will have been operated in each lane by three operators trained for that detector.

But then detector 2 is operated by three operators and the other detectors by four operators. In order to have the same number of operators for each detector, it may be necessary to remove all data set for one operator trained for detector 1 (say, operator A, in green in the table) and one operator for detector 3 (say, operator I, in blue in the table). Then after both rounds are executed, each detector will have been operated in each lane by three operators trained for that detector. And this reduced data set is as balanced as the original.

Then for one missing data set, 17 others (more than 20% of the available data) should be deleted.

3.8. Should the test be in controlled or realistic conditions environment?

There is a test and evaluation protocol to test mechanical equipment for mine action: T&E Protocol 15044 [11]. These guidelines explain how to create test lanes. Test centres exist with premises that permit testing machines based on this document.

The advantage of following such guidelines is that tests can be reproduced. They form reference to compare machines.

But the conditions of these tests might be different from conditions encountered in real mines areas. The type of soils, the vegetation, the slope, etc. may differ in reality from the test conditions. It has been argued that some machines could succeed in test sites and fail in reality in some situation. Therefore, some people prefer trials to take place in a mined country, close to real mined areas and close to where the machines will be actually used.

These field tests have also some drawbacks. The environmental conditions may vary from one place to the next, and the same equipment may have to be tested several times in different locations of the same country.

3.9. Which data to show and how to display them?

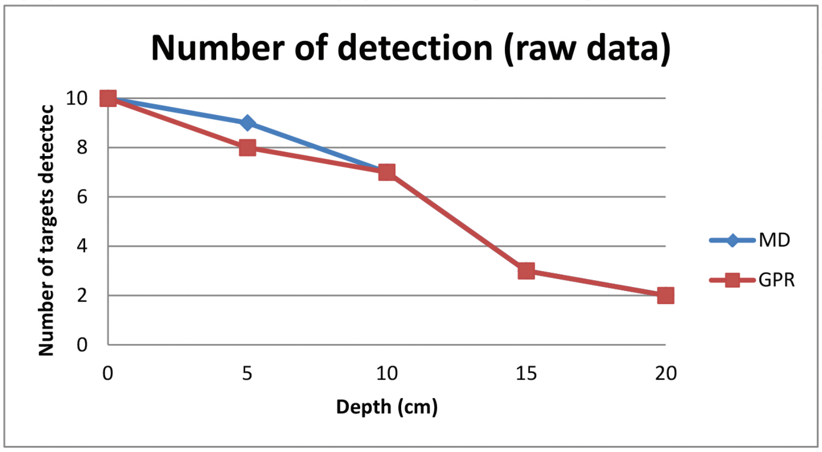
Ideally, all data should be reported and raw data should be displayed. But if there are a lot of data or if the data are noisy, some filtering may be used. Usually, models are used to filter data and fit a line, or any other curve, to them. These mathematical tools must be used carefully because they may alter how results are perceived.

Here is an example based on a real case. The test is designed to evaluate the performance of a mine detector combining a metal detector used as primary detector and ground-penetrating radar used as a confirmation tool. The metal detector is first used and when there is an alarm the ground-penetrating radar is switched on. That way, the ground-penetrating radar can only reduce the number of false alarms but cannot improve the number of detections.

Let us consider an experiment where the detector must detect targets buried at 0, 5, 10, 15 and 20 cm and where 10 targets are buried at each depth. Now let us assume that the results of the test were as described in **Table 3** and illustrated in **Scheme 3** where the ground-penetrating radar confirms all alarms from the metal detector but one at 5 cm depth.

Depth (cm)	0	5	10	15	20
Nb of targets detected by the metal detector	10	9	7	3	2
Nb of targets confirmed by the GPR	10	8	7	3	2

Table 3. Example of detection results of a metal detector and a GPR.



Scheme 3. Example of raw data.

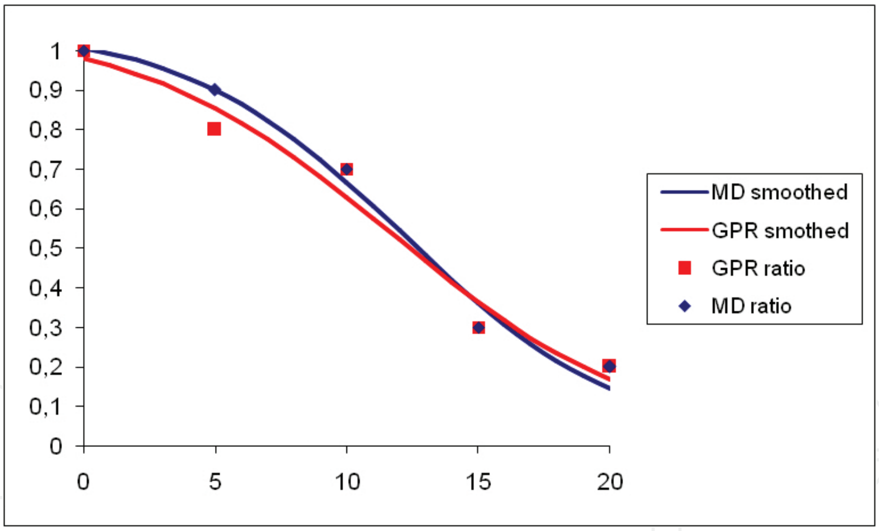


Figure 2. Example of discrete data wrongly displayed as continuous curves.

In order to display the results, let us smooth the data to fit the following curve to them:

$$y = \frac{a}{1 + e^{bx/c}}. \tag{1}$$

where *a*, *b* and *c* are parameters to estimate by least-mean square.

The results can be seen in **Figure 2**. The GPR curve goes above the metal detector curve for depths larger than 15 cm. The raw data are correct, but the graph is not.

Each point on the curve depends on the number of targets used. If confidence intervals for the probabilities of detection were displayed on all points of the curve it would make it clear that the difference between the two curves were within the confidence intervals.

Figure 3 assumes a number of targets per depth and displays the confidence interval as a function of the hit ratio. The dashed line is for five targets and the solid line for 10 targets.

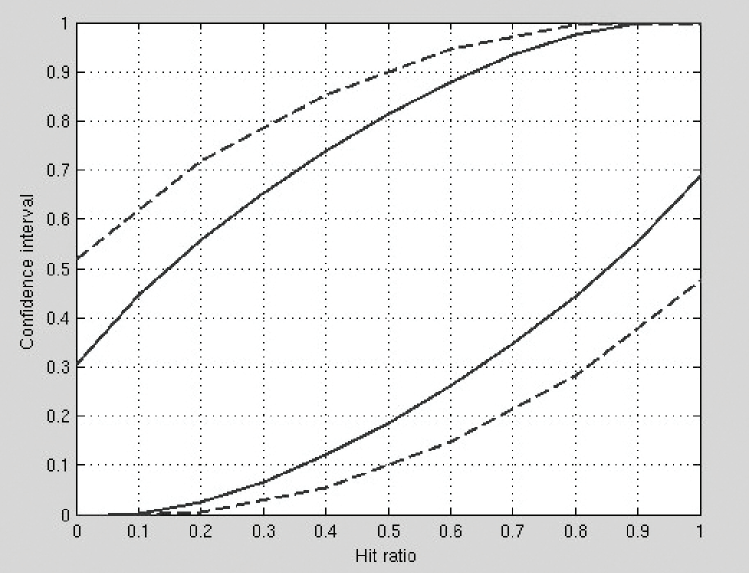


Figure 3. Confidence intervals as a function of the hit ratio for 5 targets (dashed line) and 10 targets (solid line).

For instance, if 20% of the targets are detected (hit ratio of 0.2) the confidence interval is (0.01, 0.72) for five targets and (0.03, 0.56) for 10 targets. The small difference between the results of the metal detector and the ground-penetrating radar for the depth of 20 cm (both with a hit ratio around 0.2 for 10 targets) in the previous graph is clearly not significant compare to the (0.03, 0.56) confidence interval.

Another solution would be to assume a model for the probability of detection as a function of depth and fit it to the data.

4. The case of ADE 651

4.1. Introduction

On 23 April 2013, Jim McCormick was convicted on fraud in the United Kingdom. He had sold a device called ADE 651 that, he said, could detect tiny amount of explosive from a great distance through any barrier or obstacle, but was absolutely inoperative [6].

None of the explanation given by Jim McCormick’s company, ATSC or any vendor, made any sense and all tests of the device proved that it did not work better than random chance.

In 2010 the Royal Military Academy had the opportunity to evaluate the performance of a version of ADE 651 designed, according to their producers, to detect both narcotics and explosives. This test, which proved the inefficiency of the device, led to a testimony during Jim McCormick's trial in 2013. The protocol that was used could be seen as an example of an entry test [12].

The test protocol was based on some basic principles.

4.2. Testing in optimal conditions

Since there were legitimate concerns about the very principle of ADE 651 the objective of the trial was to evaluate the performance of the device in optimal conditions.

The substance to detect would be present in sufficient quantity for detection, as agreed upon with an experienced or trained operator. Prior to the trial the operator could check whether the quantity could be detected.

The location of the trial was chosen so that there were no unwanted alarms prior to the trial. No obstacles would be placed between the substance and the operator except a screen to hide the presence or absence of substance. For instance the substance to test could be hidden in an opaque bag and case.

The device should be calibrated, if needed, to be able to detect the sub-stance.

There were no time limits for detection.

4.3. Double-blind evaluation

The test protocol requires at least three persons. First, a trained operator will operate the device under test. The samples containing either the substance to detect or something that cannot be detected would be prepared by a second person, called the sample provider, who will have no contact whatsoever with the operator. A third person, called the handler, will take the samples one at a time without knowing if they contain the substance of not, take them to the agreed-upon location and after the operator has made a determination will tell it to the sample provider. Only the sample provider will know if any given sample contains the substance to detect or not and will not have any direct contact with the operator. Moreover, the operator will have contact only with the handler, who will not know the contents of the sample.

4.4. Statistical relevance

The content of each sample, the substance to detect or not, is selected trough a random process. The sample provider prepares the sample so that neither the handler not the operator can guess whether it contains the substance or not. Then the handler brings the sample to a specific location. The operator then uses the device to determine whether the sample contains the substance. This process is repeated with different randomly prepared samples.

When testing the ADE 651 in 2010 the Royal Military Academy used a quantity of cannabis of 7–10 g. The trials used 66 samples chosen so that each sample had a uniform probability of 0.5 to contain the substance. This random selection led to 26 samples containing cannabis and 40 containing small pieces of paper.

The results of the tests are given in **Table 4**.

		Reality		
		Cannabis	**No cannabis**	
Claims	**Drugs**	10	14	**Positive predictive value 42%**
	No drugs	16	26	**Negative predictive value 62%**
		Sensitivity 38%	**Specificity 65%**	

Table 4. Results of a double-blind test of ADE 651.

We can draw the following conclusions:

- When cannabis was present it was successfully detected 10 times out of 26 (detection rate, also called true positive rate or sensitivity, of 38%)
- When cannabis presence was claimed it was true 10 times out of 24 (positive predictive value of 42%)

Other ways to interpret the results are the following:

- False positive rate or type I error is 35%
- False negative rate or type II error is 62%

The poor sensitivity and positive predictive value obtained in optimal conditions showed no raison to consider this device for further testing. A more complete statistical analysis of these results can be found in Ref. [12].

5. Evaluating external effects on performance

The effect of external conditions on the performance of a device can be important. It is therefore paramount to take that into account during a test, either by considering several different conditions or by selecting a set of conditions and evaluating the equipment during these conditions.

An alternative way to deal with this problem is to evaluate the effect of some conditions on the performance. An easy way could be to measure proxies to this effect. The Royal Military Academy worked in this direction when it co-ordinated the work of drafting a test protocol to evaluate the effects of soils on the performance of metal detectors and ground-penetrating radars.

This work was performed during a CEN Workshop chaired by the Royal Military Academy [13]. This Workshop produced a document, T&E Protocol 14747-2, which provides some information about how soil may affect the performance of metal detectors and ground-penetrating radars and guidelines to estimate these effects and the soil characteristics that cause them [14].

It first gives guidelines that cover measurements that can be done during field operations. These are easy to follow and give a rough indication of the effect of soils on performance.

It also gives a list of characteristics to measure and methods to measure them. This includes magnetic soil properties such as susceptibility, effective relative electric permittivity, effective electrical conductivity, attenuation coefficient, characteristic impedance, the surface roughness or soil water content, characteristic of targets such as its electric size, how to document weather conditions, soil texture, how to describe vegetation, roots, rocks, surface cracks, etc.

6. Existing test protocols for mine action

The European Committee for Standardisation (CEN) has published several CEN Workshop Agreements containing test protocols for mine action. They are now called T&E Protocols and are available together with IMAS.

T&E Protocol 14747-1:2003, formerly known as CWA 14747-1:2003, provides protocols to test and evaluate metal detectors [8].

T&E Protocol 14747-2:2008, formerly known as CWA 14747-2:2008, focuses on characterising the effect of soil on the performance of metal detector and ground-penetrating radars [14].

T&E Protocol 15044:2009, formerly known as CWA 15044:2009, provides protocols to test demining machines [11].

In addition, many reports of test campaigns carried out under the ITEP programme are publicly available on the GICHD website.

7. Conclusion

Is a new technology better than current technologies? How do we define ‘better’ equipment? Should I purchase a given tool or another? What are the performance and limitations of a given solution? Trials are a method to provide answers to such questions. It is therefore important to clarify the questions and to design a test protocol that will permit to gather evidence to provide the answers. The importance of test protocols therefore should not be underestimated. In case poor protocols are used no reliable conclusions might be drawn and the trials might be a waste of time and resources and their conclusions disputed.

Test protocols are an integrant part of a test report. A test report cannot be correctly analysed and no valid conclusions may be drawn from it without an understanding of the protocol that was used.

Designing a good test protocol is not easy. This chapter gives a list of things to keep in mind, based on the long experience of the Royal Military Academy.

Fortunately many protocols, test reports and guidelines have already been published. These publications may be good starting points when designing your own protocol.

Moreover, published test results may already give you the answers you are looking for without having to carry out trials yourself.

Acknowledgements

The author of this chapter would like to thank the Belgian Ministry of Defence and the European Union.

The trial of ADE 651 was carried out with Pascal Druyts (RMA) and the support of the Police Zone of Geel-Laakdal-Meerhout.

Author details

Yann Yvinec

Address all correspondence to: yann.yvinec@rma.ac.be

Royal Military Academy, Department Communication, Information, Systems & Sensors,
Brussels, Belgium

References

[1] Enright J.T., “Testing dowsing: the failure of the munich experiments”, Skeptical Inquirer 23(1), January/February 1999. http://www.csicop.org/si/show/testing_dowsing_the_failure_of_the_munich_experiments, accessed on 8 September 2016

[2] Ioannidis J.P.A., “Why most published research findings are false”, PLoS Med 2(8), e124, doi:10.1371/journal.pmed.00201242005, Published 30 August 2005

[3] Button K.S., Ioannidis J.P.A., Mokrysz C., Nosek B.A., Flint J., Robinson E.S.J. and Munafo M.R., “Power failure: why small sample size undermines the reliability of

neuroscience", Nature Reviews Neuroscience 14, 364–376, May 2013, doi:10.10.1038/nrn3475

[4] Johnson V.E., "Revised standards for statistical evidence", Proceedings of the National Academy of Sciences of the United States of America, 19313–19317, doi:10.1073/pnas.1313476110, https://www.researchgate.net/publication/258446201_Revised_Standards_for_Statistical_Evidence, accessed 8 September 2016

[5] Australian Research Centre for Health of Women and Babies, "Effectiveness of Homeopathy for Clinical Conditions: Evaluation of the Evidence", Overview Report Prepared for the NHMRC Homeopathy Working Committee by Optum October 2013. https://www.hri-research.org/wp-content/uploads/2016/02/Literature_review_of_public_submissions.pdf, accessed on 8 September 2016

[6] BBC, "Fake bomb detector seller James McCormick jailed", 2 May 2013, http://www.bbc.com/news/uk-22380368, accessed on 14 January 2016.

[7] Gülle D., Gaal M., Bertovic M., Müller C., Scharmach M., Pavlovic M., "South-East Europe Interim Report Field Trial Croatia—Continuation of the ITEP-Project Systematic Test and Evaluation of Metal Detectors (STEMD)" available at www.gichd.org/fileadmin/pdf/LIMA/STEMETAL DETECTOR_Interim_Croatia_final.pdf, accessed on 31 January 2013.

[8] "T&E Protocol 14747-1:2003", formerly known as CWA 14747-1:2003, available at www.mineactionstandards.org

[9] Takahashi K. and Gülle D., "ITEP evaluation of metal detectors and dual-sensor detectors", Journal of ERW and Mine Action, 14(3), pages 76-79, Centre of for International Stabilization and Recovery at James Madison University, Fall 2010

[10] "Guidelines for Reliability Tests of Dual Sensors in Humanitarian Demining", 21 June 2010, http://www.gichd.org/fileadmin/pdf/LIMA/DSTESTguidelines2010.pdf, accessed on 15 January 2016, pages 76 to 79

[11] "T&E Protocol 15044:2009", formerly known as CWA 15044:2009, available at www.mineactionstandards.org

[12] Yvinec Y. and Druyts P., "A Simple protocol for a double-blind test on an explosives/drugs long-range detector" in International Conference dedicated to Hazardous Materials: Issues of Detection and Disposal, Koscierzyna, Poland, May 2010.

[13] Druyts P., Yvinec Y., Acheroy M., "Relating soil properties to performance of metal detectors and ground penetrating radars", in Using Robots in Hazardous Environments, Elsevier, 2011, Copyright © 2011 Woodhead Publishing Limited. All rights reserved.

[14] "T&E Protocol 14747-2:2008", formerly known as CWA 14747-2:2008, available at www.mineactionstandards.org