Proposal for a Standard for Reliability Tests of Dual Sensors in Humanitarian Demining

Draft 01, 12 April 2007

Mate Gaal, Dieter Guelle, Marija Bertovic
Federal Institute for Materials Research and Testing
(Bundesanstalt für Materialforschung und –prüfung, BAM)
Berlin, Germany

Foreword

This document is a proposal for standardised reliability tests of dual sensors for humanitarian demining. It is conceived as a fragment of a possible dual sensor CEN workshop agreement, which would provide guidelines, principles and procedures for the testing and evaluation of dual sensors. To a large extent this document is based on the standard for testing metal detectors for humanitarian demining CWA 14747:2003 [1]. CWA 14747:2003 was approved by CEN and verified in practice. A number of trials has been performed in recent years based on that standard. BAM (German Federal Institute for Materials Research and Testing) organised a series of four trials in Germany and Croatia [2][3][4]. The STEMD (Systematic Test and Evaluation of Metal Detectors) project organised by JRC (Joint Research Centre of the European Commission, Ispra, Italy) comprised field tests in Laos and Mozambique, and a comprehensive lab test [5][6][7][8]. A trial in Croatia was organised by BAM as a continuation of the STEMD project [9][10]. The Lao and the Croatian trial included reliability tests. Experiences from dual sensor trials performed in the recent years [11][12][13][14] and previous articles about dual sensor testing procedures [15][16] are also built in this document.

The chapters of the CWA for metal detectors include: 1 to 5 the general conditions for testing and reporting, 6 tests in air to different targets, 7 environmental conditions and their influence, 8 in soil detection capability (including the reliability trial), 9 operational capabilities (pinpointing, adjacent targets etc.), and 10 ergonomics and ruggedness. The annexes give details for soil characterisation, test targets (there are no yet such targets for dual sensors), and reporting forms. 31 test is described in that standard and there are instructions on how to report the results.

A CWA for testing dual sensors would include several different tests. Reliability tests described in this document will be one chapter or a section in that standard, similarly as the stand-alone metal detector reliability tests are described in Section 8.5 of CWA 14747:2003. For that reason this document is not meant to be final; it will be connected with references to other tests. The CWA for soil may serve as an example that there are activities going on closely connected with the standardisation of test procedures. It will be a second part for the CEN Workshop Agreement on Test and Evaluation of Metal Detectors specifically on soil characterisation. This standard will define soil factors influencing the metal and dual sensor performance, which are not a subject of the present proposal. (RMA /Royal Military Academy/ Belgium is in charge of working out the Part 2 of the CWA 14747 concerning soil characterisation and all proposals with that respect should be directed to RMA.)

The following is proposed for a public discussion with the aim of reaching an agreed document:

- This document will be published at the BAM website. The internet address will be sent to the interested parties as soon as it is available on internet.
- The document will be open for a public discussion. It will be updated monthly and the latest version of the document will be available on the above-mentioned web-site. The aim is to finish the discussion and have an agreed document 16 July 2007.
- The distribution list will include ITEP WGMS (International Test and Evaluation Programme for Humanitarian Demining, Working Group Multi-sensors), UNMAS (United Nations Mine Action Service), GICHD (Geneva International Centre for Humanitarian Demining), manufacturers of demining equipment and persons recommended by them, mine action organisations and centres recommended by UNMAS as well as other interested persons.
- Every addressed participant or other interested person can make comments and proposals for the improvement of the document. This information will be collected, analysed and included where appropriate. If such proposals are refused or directed to other places, an explanation will be given by the authors.
- The exchange of information should go via the distribution list. Comments and proposals can be sent to the distribution list or directly to Mate Gaal (mate.gaal@bam.de) and Dieter Guelle (dieter.guelle@bam.de).

In the IMAS series of standards, the words 'shall', 'should' and 'may' are used to indicate the intended degree of compliance. This use is consistent with the language used in ISO standards and guidelines:

a) 'shall' is used to indicate requirements, methods or specifications that are to be applied in order to conform to the standard;
b) 'should' is used to indicate the preferred requirements, methods or specifications; and
c) 'may' is used to indicate a possible method or course of action.

CONTENT

1. Terms and definitions

All terms and symbols used in this work follow the definitions given in CWA 14747:2003. Definitions of terms and symbols which are not provided in that document are listed in this and the next chapter.

dual sensor
A device which is a combination of a metal detector and a ground penetrating radar.

false alarm rate
The estimated false alarm rate is defined as the number of false alarms counted on an area divided by the size of that area, or the average number of false alarms per square metre. The area is calculated as the area of the test lane minus the area of all detection halos. Measurement unit: $1/m^2$.

false positive indication
False alarm, i.e. a positive indication of an operator when a target is not present.

ground penetrating radar
A device which uses electromagnetic radiation in the microwave band to detect the change in the dielectric constant in the ground.

metal detector
The component of the dual sensor which uses the principles of electromagnetic induction to reveal the presence of metal in its vicinity.

stand-alone metal detector
A stand-alone device which uses the principles of electromagnetic induction to reveal the presence of metal in its vicinity. NOTE: This term corresponds to the definition of metal detector in CWA 14747:2003.

operator
A person operating a dual sensor during a test.

POD curve
A curve of POD in dependence on depth.

probability of detection
The estimated probability of detection for a particular combination of an operator, dual sensor and a lane is the ratio of the number of detected targets and the total number of opportunities to detect a target.

sensitivity area
Sensitivity area or sensitivity cone is the area below the search head within which a particular target causes the detector to alarm.

sensitivity cone
see “sensitivity area”

sensitivity profile
A projection of the sensitivity area to a plane. NOTE: Sensitivity area is three-dimensional, sensitivity profile is two-dimensional. The plane of highest relevance is the plane separating the dual sensor symmetrically into the left and the right halve. The sensitivity profile in that plane determines the safe advance of the search-head.

targets
see “test targets”

test targets
Test targets (or targets) are objects used to test the detection performance of both components of the dual sensor. They may be inert mines, mine simulants or any objects representing the objects that need to be detected.

ROC diagram
A diagram of POD versus FAR.

run
A complete pass of an operator with a dual sensor through a lane.

start
A group of runs performed at the same time.

test objects
Objects deliberately buried in test lanes. In dual sensor reliability tests, there are two kinds of test objects: test targets (or targets) and metal clutter.

true positive indication
A positive indication of an operator if a target is present.

2. Abbreviations and symbols

DS
dual sensor

FAR
false alarm rate

FP
false positive indication

GPR
ground penetrating radar

POD
probability of detection

ROC
receiver operating characteristic

TP
true positive indication

3. Detection reliability tests

3.1 Principle

The objective of detection reliability tests is to evaluate the detection reliability of the mine detection device when used by an operator who does not know the location of the targets. Detection reliability is the degree to which the mine detection device is capable of achieving its purpose, which is to have maximum capability for giving true alarm indications without producing false alarm indications. Detection reliability tests are statistical performance tests in which the effect of environmental and human factors is added to the intrinsic capability of the devices. Detection reliability tests are typically performed in or near to an area to be cleared of mines, in representative local soil and with targets representative of the local mine threat. Such tests may be part of either "consumer report" trials or "acceptance trials".

The objects deliberately buried in the test lanes are called test objects. There are two kinds of test objects: test targets and metal clutter. Test targets (or targets) are objects used to test the detection performance of both components of the dual sensor. The metal clutter is used to test the ability of the GPR to identify the MD signals coming from clutter as false alarms.

There are three ways dual sensors confirm mine-like targets. [1] These are:

a) the operator switches the GPR on manually when he receives a signal from the metal detector,
b) the GPR is switched on automatically whenever the metal detector produces an alarm tone,
c) the operator receives a processed signal as a result of data fusion for a confirmation of a mine-like target.

---

[1] At their current stage of development, dual sensors still do not provide a data fusion of both sensor signals. However, at least one manufacturer is already working on producing a data fusion dual sensor. It is therefore of value to include data fusion dual sensors in this document.

In dual sensors of type a) and b) the primary sensor is the metal detector: only after the metal detector gives a signal, the GPR is switched on. If the signal would be confirmed by the GPR, it would be investigated in a minefield. The testing procedure of dual sensors of type a) and b) is described in the following lines. The operators use coloured markers of two different colours to mark the locations of MD alarms and the GPR alarms. The positions of their indications are measured and recorded. They are compared with the actual positions of test targets and the so called detection halo radius, which is defined in Annex A. If a marker lies within the detection halo, the target is detected and the indication is called a true positive indication. Markers outside the halo radii of test targets are counted as false alarms and called false positive indications. Markers in proximity of metal clutter are also counted as false alarms. Since two kinds of markers are used, the performance of the MD and of the GPR can be evaluated separately. A device with high detection reliability is one that maximizes the number of true alarm indications and minimizes the number of false alarm indications.

In tests of dual sensors type c) (data fusion dual sensors) markers of only one sort shall be used. In other respects the testing procedure shall be the same as for other dual sensors.

3.2 Lanes

A site shall be chosen for the tests at which:

- the native soil at the test site is representative of an area to be cleared of mines (the test site may be within that area), or
- soil representative of the region requiring mine clearance has been placed in test lanes.

Note that if soil is moved, the dual sensor response will not necessarily be the same as it was prior to being disturbed. The soil in which it is most difficult to detect mines (the noisiest) found in the mine-affected area should be used where possible. Either the test site should be set up in an area of that soil, or that soil should be all be transported and used in a test lane.

The vegetation over lanes shall be removed or cut short (so as not to impede sweeping as close as possible to the ground[2]). Any metal objects on the soil surface shall be removed. Following this, using a stand-alone metal detector, remove any buried metal objects. In some cases, during this clearing procedure, detection signals may come from the soil itself and be mistaken for metal. Therefore, some judgement will have to be exercised in determining how much effort is spent on metal object removal to minimize alarm indications from sources other than the placed targets. The goal is a test site free of metal objects other than the targets placed intentionally, so that detection of mines can be assessed.

---

[2] Annex B.5 of CWA 14747:2003 specifies: “Normal maximum sweep height above the surface of the ground shall be 30 mm.” The specification of the maximum sweep height is redundant, because it is not possible to monitor the sweep height during a reliability test.

If soil is brought in from another area for construction of the test site, the soil should be compacted so that it resembles as far as practicable the state found in an actual demining area – although this does not guarantee to give the same response of a dual sensor as at the original place.

The test areas used for blind detection tests shall comply with the requirements of Section 8.1.2 of CWA 14747:2003 as a minimum. However, in order to perform such testing effectively a large area is needed, which should be divided into well defined test lanes. Corner locations of the test lanes can be marked with non-metallic corner posts or pegs. These stationary markers serve to define the test lane boundaries and are used as the reference points for measuring actual target locations and locations at which target detections are declared. Lane width should be 1 m and depth at least 0,5 m. Two stripes at least 25 cm wide shall be established on both sides of the lane. The lane and the two stripes shall comply with the requirements of Section 8.1.2 of CWA 14747:2003 mentioned above.[3] The width and depth of the lanes are so specified to ensure there is no effect from the indigenous soil of the test site (if different from the soil in the test lane) on the dual sensor during the test.

A training and calibration area should be established where operators can practise using the dual sensors assigned to them for the test. The area should include a metal-free area as specified in 8.1.2 of CWA 14747:2003. In addition there should be a similar area containing targets as used in the test lanes. This area should be located away from the test lanes to minimize the possibility of operators picking up clues as to where mines are buried in the test lanes to be used for blind tests.

An accurate method is required to measure and record the placement of test objects and to record the location of detections declared during the field trials. A laser-based total station survey system is ideal for this purpose. Three or four reference benchmarks outside the lanes should be measured immediately before and after each set of lane corner, target, or detection-marker measurements to confirm data integrity and provide some chance of data recovery due to operator error. If such a survey system is not available, 100-m non-stretch tape measures should be provided to measure these positions relative to the reference point. These tape measures are also required to facilitate the laying out of the test site.

Once the test objects are laid in the lanes, all visual clues that might indicate a presence of a test object shall be removed. Sweeping or lightly raking the test lane surface can remove such clues. If tests extend over a period of days, lanes shall be inspected daily to remove clues that may result from settling, rain, and other causes. Even with these safeguards, telltale clues can still occur. Therefore, operators shall be told that the test is to measure detection reliability of the detector (and is not a test of individual operator performance) and therefore to ignore any visual evidence of target location.

---

[3] In the CWA 14747:2003, the lanes are defined to be 1,5 to 2 m wide and the targets are placed in the 1-m stripe in the middle of the lane. The formulation provided in this document is simpler and corresponds to the meanwhile established use of the term “test lane”.

Electromagnetic and other soil properties influencing the dual sensor performance shall be measured as defined in CWA 14747 Part 2[4].

3.3 Test objects

Two kinds of test objects shall be used: test targets and metal clutter. Test targets (or targets) are objects used to test the detection performance of both components of the dual sensor. Test targets may be inert mines, mine simulants or any objects creating a similar signal as the objects that need to be detected. Metal clutter is used to test the ability of the GPR to identify the MD signals coming from clutter as false alarms. The clutter needs to be representative of the region of interest.

Since it is difficult to acquire inert mines or even their faithful surrogates, the trial organiser may be forced to use some kind of standard test targets. Such targets (known as “ITOP inserts”) exist for stand-alone metal detectors and are mentioned in the standard CWA 14747:2003. Internationally accepted standard targets for dual sensors or even only for GPR still do not exist.

3.4 Layout, depth, orientation and separation of test objects

Test objects shall be placed to random locations within the lane, with their entire halo inside the lane. The test objects and their depths shall be intermixed randomly down the length of the lane. Separation of test objects in the test lane shall be large enough to ensure that the device under test cannot give an alarm indication due to the response from two test objects at the same time. The separation shall be at least 0,5 m. This condition limits the number of test objects in a lane. The number of test objects in a lane should be approximately between 1 and 1,2 per square metre (for example, 30-36 test objects in a 30-m lane).[5] In each lane there shall be at least one area 1 m long without any test objects. The number of clutter pieces should be roughly equal to the number of test targets.

The orientation of buried mine targets shall be as normally laid – horizontal with activation device uppermost.

It would be pointless to burry the targets for a blind reliability test to a depth at which they cannot be detected in an open test. The test targets of a certain type should be placed to depths between zero and the expected average maximum detection depth for that target

---

[4] The work on this standard for soil characterisation for metal detector and ground penetrating radar (GPR) performance evaluation is in progress. It is lead by the Royal Military Academy (RMA) in Brussels, Belgium.

[5] It is not recommended to place more than 1,2 test objects per square metre in a lane because that would lead to formation of patterns.

type measured with the most sensitive dual sensor in the test and established in repeated measurements.[6] In most cases, repeated measurements of maximum detection depth are not available prior to trial; in that case the average maximum detection depth needs to be estimated from the available knowledge. The metal clutter shall be buried to such small depths, that it is very easily detectable by a metal detector.

The choice of target depths depends on the analysis that will be performed after the test. The simpler option is to bury the targets at one or more of the depths specified in CWA 14747:2003 (flush with the surface, 5cm, 10cm, 15cm and 20cm), depending on local clearance depth requirements, the mine types of interest and the results of in-soil maximum detection depth measurements.[7] Several pieces of the same mine type would be buried to the same depth. The results are evaluated for each depth separately. The second option is to bury the targets to depths in smaller steps, for example, 1, 2, 3, 4 ... cm, only one or two targets to each depth. The corresponding data evaluation is described in Section 13].

As mines are buried, their location shall be accurately noted. The location shall be expressed as down-lane and cross-lane measurements from the end and sides of the test lane (using the corner posts or pegs as datum points). The measurement of each target actual location, its type, and its depth, shall be recorded for each test lane. Access to this information shall be restricted to personnel managing the test and shall not be disclosed to any of the operators participating in the test.

3.5 Operators

The operators shall be capable of using the dual sensor as intended by the manufacturer and they shall be representative of the operators that would use the dual sensor in the field. The operators shall be currently active deminers, since they represent the persons who will actually perform clearance operations. They should have at least 6 months of continuous work in minefields directly before the trial. The goal of this CWA is to provide an objective evaluation of the capabilities of the dual sensor. However, since an essential part of the dual sensor operation is interpretation of alarm indications by the operator, variable human factors are introduced. In order to minimize the effect of a single operator on the results, at least three (3) operators shall use each dual sensor type. The recommended number of operators is at least six (6). If possible, all of the operators should operate each dual sensor in each test lane. If not, groups of operators should be created for each dual sensor. It should be ensured that these groups have similar characteristics (age, demining experience). If necessary, a qualification performance test

---

[6] CWA 14747:2003 prescribes: “Testing with targets at depths of 50 and 100 mm more than those derived from the in-air maximum detection height should ensure that the depth range used spans the maximum detection depth.” However, it is clearly better to conform to the maximum detection depth measurements, if available and if they include repeated measurements.

[7] CWA 14747:2003 specifies: “For each target type buried, at least seven (7) identical targets shall be buried at the same depth. At least 28 targets shall be used in each test soil.” This specification is not necessary.

can be made so that the actual performance of an operator in that specific test site can also be estimated and controlled.

Operators should be trained on the proper use of each dual sensor that they will operate during the reliability test. They should have adequate time to become familiar with the operation of the dual sensor, assure themselves that each dual sensor is able to detect the test targets in the calibration area and to attune their ears to the sounds of the device. They must feel no time pressure. At least two weeks of training for each dual sensor model are recommended.

Some elements of the local standard operating procedures and the local demining practice may be applied to improve the concentration of the operators and to make their work similar to their daily routine. The most important element may be the presence of a section leader performing quality assurance.

NOTE: Deminers in any given demining operation are frequently familiar with specific stand-alone metal detectors. This familiarity could affect the results of a test. Therefore, the make and model of the metal detector they have been using shall be noted.

3.6 Test procedures

At the start of each test, the dual sensor shall be set up according to the manufacturers instructions as given in the operation manual. The dual sensor shall be adjusted for the maximum sensitivity attainable on the soil of the test lane. The metal detector shall be set up to the soil if necessary and the GPR calibrated in the calibration area before each run. The detection tests shall be blind; that is to say the operator shall not know the location of targets laid in the lane prior to searching for them with the dual sensor. The operator shall be supervised at all times by personnel managing the testing. Each dual sensor shall be used in its normal operating mode, sweeping manually in a transverse direction at a speed that ensures optimum detection capability (see Sections 6.4.2 and 6.4.3 of CWA 14747:2003). The operator shall move the dual sensor forward along the test lane between sweeps. The distance in the forward direction between successive sweeps shall be small enough to ensure that the sensitivity area of both sensors covers all of the ground in accordance with the target depth. The appropriate sweep overlap may be determined by measuring the sensitivity profile according to 6.7 of CWA 14747:2003[8].

The operators shall attempt to detect all test objects in the test lane. The operators shall use coloured markers to mark the locations of MD alarms. After each MD alarm, they shall check the alarm location with the GPR. If the GPR also produces an alarm, they shall pinpoint and mark the position with a differently coloured marker, without moving the marker of the corresponding MD signal. The direction in which any single operator

---

[8] The safe advance for the MD is at the same time the safe advance for the dual sensor. The safe advance of the GPR alone is not relevant for dual sensors currently in use, since in those devices the GPR is used after the MD provides a signal and pinpoints the buried object. For data fusion dual sensors (under development) the safe advance of the MD cannot be established – the sensitivity profile and the safe advance can be established only for the device as a whole.

moves down a lane shall be alternated (bottom to top, top to bottom) with each pass. This will make it harder for the operator to memorize locations where he/she has previously detected a target. Blind tests shall not be observed by other operators for the same reason. The only observers shall be the test personnel.

Each complete pass of an operator with a dual sensor through a lane is called a run. After each run, test personnel shall measure the location of each marker, record the measurements, including the colour of the marker, and retrieve the markers. After the test, the recorded marker locations shall be compared to the actual test target locations to determine the test results. This process shall be repeated for each dual sensor in each soil type. If possible, the test shall be performed with at least two (2) examples of the same dual sensor model. When the test is over and as the targets are removed, the identity and position of the removed targets shall be verified and compared to the original location measurements. Should there be any discrepancy, the test results shall be corrected as appropriate.

If locations in the test lane give persistent false indications, such locations shall be investigated to ensure that there is no object that has been inadvertently left there. If such an object is found, all reported indications of prior tests that correspond to this object (within detection halo) shall be omitted from analysis of the results.

3.7 Design of experiment

The trial organiser needs to decide how to combine the dual sensors, operators, and lanes in the trial. Unfortunately it is not possible to give simple prescriptions applicable to every experimental problem. Each trial is different: both the available resources and the experimental goal vary from trial to trial. The experimenter should be familiar with the principles of experimental design and approach every trial as a new experimental problem. However, some general recommendations can be made.

In an ideal trial all dual sensors would be tested by all operators, in order to reduce the influence of the individual differences between the operators. However, the time required for training would than be very long. It is therefore acceptable that each dual sensor model is used by other operators.

The design of experiment shall be a fractional factorial design based on a Latin square or a Graeco-Latin square. Variable “start” shall be introduced to indicate the execution order of the runs. Examples are provided in Annex B. A single Latin square or a Graeco-Latin square shall be repeated with permuted operators or dual sensors or both, so that eventually all dual sensors will have been used by all operators trained for that dual sensor in all soil types, if the scope of the trial allows that. The number of soils and target types should be as small as possible, so that the number of targets in each target-soil combination can be as large as possible. The number of operators should be as large as possible, since operator can be understood as a nuisance factor, in the sense that the experimenter is not interested in the results of any particular operator. If some earlier measurements imply that there is very little variation between specimens (i.e. copies) of the same dual sensor model, than only one specimen may be used in a reliability test, or more specimens treated as identical.

3.8 Test results, reporting and evaluation

The output variables of a reliability test are probability of detection (POD) and false alarm rate (FAR). The estimated probability of detection for a particular combination of an operator, dual sensor and a lane is the ratio of the number of detected targets and the total number of opportunities to detect a target. This number is here denoted n.[9] The estimated false alarm rate is defined as the number of false alarms counted on an area divided by the size of that area, or the average number of false alarms per square metre. The area is calculated as the area of the test lane minus the area of all detection halos. If we assume a binomial distribution for the number of true positive indications, we can find the 95% confidence limits for the probability of detection. Similarly, if we assume Poisson distribution for the false alarms, we can construct the confidence limits for the false alarm rate [18][20][21]. They are described later in this section.

As described in section [11], detections of the metal detector and the GPR for dual sensors of type a) and b) (the GPR is switched on only after the MD gives a signal) shall be marked with markers of two different colours. The positions of the markers shall be compared with the actual positions of the test targets. The criterion for a detection is the size of the so called detection halo, which is defined in Annex A. If a marker lies within the halo radius of a test target, that target is detected and the indication is called a true positive indication. Markers outside the halo radii are counted as false alarms, also called false positive indications. Markers lying within the halo of a metal clutter piece shall be counted as false alarms.

---

[9] For example, if there are 15 test targets in a lane and 4 runs are analysed, the total number of opportunities to detect a test target is 4 times 15, that is n=60.

Since two kinds of markers shall be used for dual sensors of type a) and b), the performance of the MD and of the GPR can be evaluated separately. True positive (TP) and false positive (FP) indications shall be counted for each dual sensor component separately. The numbers of TP and FP indications for the metal detector are here marked x_{MD} and y_{MD} respectively. The numbers of TP and FP indications for the GPR are here marked x_R and y_R respectively. The number of FP for the GPR will certainly be smaller or equal to the number of FP for the metal detector ($y_R \leq y_{MD}$), because the GPR is used only after the MD makes an indication. It is almost certain that the number of TP for the GPR will be smaller than the number of TP for the metal detector ($x_R \leq x_{MD}$).[10]

	MD detections	radar detections
number of TP	x_{MD}	$x_R \leq x_{MD}$
number of FP	y_{MD}	$y_R \leq y_{MD}$

The difference $x_{MD} – x_R$ tells us about how many true positive indications of the MD the radar rejected and therewith not identified a test target. The difference $y_{MD} – y_R$ tells about how many false alarms of the MD are identified as false alarm, speeding up the clearance process.

For establishing the estimated POD, the above mentioned numbers x_{MD} and x_R shall be divided by the total number of opportunities to detect a test target (here marked n). For establishing the estimated FAR, the numbers of false negative indications shall be divided by the searched area. Thus two different PODs and two different FARs are defined: $POD_{MD}= x_{MD}/n$ and $FAR_{MD}= y_{MD}/n$ for the MD and $POD_R= x_R/n$ and $FAR_R=y_R/n$ for the radar.

The POD and the FAR shall be combined in an ROC (receiver operating characteristic) diagram, so that the FAR is on the abscissa and POD on the ordinate axis, with the corresponding 95% confidence limits. The calculation of 95% confidence limits is described in Annex C. For a particular choice of lanes, the results of each dual sensor of type a) and b) shall be represented by two points on an ROC diagram: one for the MD component, and the other for the GPR component of the dual sensor. The best achievable result for the GPR is the confirmation of all true positive indications and the rejection of all false alarms of the MD. In that case, the POD_R would be equal to POD_{MD}, while FAR_R would be zero.

FAR reduction is estimated differently for dual sensors of different types. For dual sensor of types a) and b) (dual sensors without data fusion) FAR reduction is defined as the number of false alarms indicated by the MD but identified as false alarms by the radar (y_{MD}-y_R) divided with the total number of false alarms indicated by the MD (i.e. y_{MD}).

---

[10] The number of GPR TP indications can be larger than the number of MD TP indications only if pinpointing is much more precise with the GPR than with the MD. Than it could happen that some MD indications fall just outside the detection halo, while the GPR indications around the same targets fall inside the halo. However, this is most unlikely to happen.

FAR reduction is therefore $(y_{MD}-y_R)/y_{MD}$. It can also be calculated as $1 – FAR_R/FAR_{MD}$, which gives the same result.[11] For example, if two runs were performed, and if the number of false alarms by a metal detector was $y_{MD} = 12$, and the number of false alarms by a GPR was $y_R = 3$, than the FAR reduction is (12-3)/12 = 0,75 = 75%.

The indications of dual sensors of type c) (data fusion dual sensors) shall be marked with markers of one colour. In all other respects, the testing procedure shall be the same as for other dual sensors. The resulting ROC diagram shall contain only one point. The ability to deal with MD false alarms shall be evaluated by analysing the indications around metal clutter. If a marker is placed within the detection halo of a metal clutter piece, the piece is considered to be indicated. The estimated false alarm rate reduction is defined as the number of metal clutter pieces which are not indicated divided with the total number of opportunities to detect a metal clutter piece. For example, if only one run is analysed, if there are 10 clutter pieces in a lane and if 3 of them were indicated, than the estimated false alarm reduction is 7/10 = 0,7 = 70%. An ideal dual sensor would produce no indications around metal clutter pieces and the FAR reduction would be equal to 100%.

The other kind of diagrams used in reporting shall be POD curves. POD curves are curves of POD in dependence on target depth. For creating POD curves, the best choice is to bury the test targets to depths in smaller steps, for example, 1, 2, 3, 4 ... cm, only one or two targets to each depth. The creation of POD curves is described in Annex D.

A simpler analysis of POD in dependence on depth is possible if targets are buried to a smaller number of depths, for example, flush with the surface, 5, 10 and 15 cm deep. Detections for each depth are simply counted separately and expressed as POD with the appropriate confidence limits, in a diagram of POD versus depth [2][4]. The confidence limits are calculated with the assumption that the POD is binomially distributed. Such a method has the advantage that it needs fewer calculations and that it does not depend on any assumptions about the relationship between the POD and the depth; the only assumption is that the detections on a certain depth are binomially distributed. Its disadvantage is that it produces larger than necessary confidence intervals and that it cannot give information about the depths not present in the test.

---

[11] Some experts [13] proposed to examine only the false alarms caused by the metal clutter, but there is no reason to restrain the analysis on metal clutter, unless with data fusion dual sensors. The purpose of dual sensor is not to “filter our” clutter, but false alarms of all kinds, for example, caused by soil or unknown metal pieces.

Annex A: Detection halo

During blind tests, the locations at which operators report alarm indications shall be recorded. To determine whether these indications can be considered to be from the intended targets (i.e. they are true indications) or are false indications, the distance of the indication location from actual target locations shall be used. If the indication location is within a certain distance of a target location it shall be considered a true detection indication. If the indication is further than this from the target location, it shall be considered a false detection indication. If there are several indications within the halo, all accept one of them shall be ignored. The circle around the target location whose radius is defined by this maximum distance is known as the detection halo. The radius of the detection halo shall be half of the maximum horizontal extent of the target plus 100mm.

Annex B: Design of experiment based on a Latin square

Table 1 is an example of a 4x4 Latin square. Table 2 is an example of a 4x4 Graeco-Latin square. These squares are recommended as the basis for a trial design if four dual sensors are to be tested. Dual sensors are here marked with letters A, B, C and D. In Table 2, the letters mark the operators and the Greek letters mark the dual sensors. For other number of dual sensors, other Latin and Graeco-Latin squares should be used. Latin squares and Graeco-Latin squares are described in [17][18][19] and their use in metal detector tests in [1] and especially in [12]. Graeco-Latin squares should be used if all operators are trained for all devices. In that case, crossover design should be applied [12]. With crossover design, it is possible to achieve that the operators work with fewer detector models at a time. This is strongly recommended, since it reduces stress on the operators.

Table 3 is an example of an experimental design based on a 3x3 Latin square. The variable “start” marks the order of execution. The experimental problem is to test three detector models, two specimens each, in six lanes. The letters A-L denote the operators, while the numbers denote the dual sensors. The first number (1, 2 and 3) denotes the dual sensor model, the second number (1 and 2) the specimen (i.e. the copy) of the same model. Persons A B, C and D are trained to work with dual sensor 1; persons E, F, G and H with dual sensor 3; and persons I, J, K and L with dual sensor 3. The detector models form 3x3 Latin squares; four of these squares are combined to form a round. After both rounds are executed, each detector will have been operated in each lane by all four persons trained for that detector.

Table 1. Example of a 4x4 Latin square.

	start 1	start 2	start 3	start 4
lane 1	A	B	C	D
lane 2	B	A	D	C
lane 3	C	D	A	B
lane 4	D	C	B	A

Table 2. Example of a 4x4 Graeco-Latin square.

	start 1	start 2	start 3	start 4
lane 1	A alpha	B beta	C delta	D gamma
lane 2	B gamma	A delta	D beta	C alpha
lane 3	C beta	D alpha	A gamma	B delta
lane 4	D delta	C gamma	B alpha	A beta

Table 3. Example of an experimental design based on a 3x3 Latin square.

		start 1		start 2		start 3		start 4		start 5		start 6	
round 1	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

		start 1		start 2		start 3		start 4		start 5		start 6	
round 2	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

Annex C: Confidence limits for the POD and the FAR

The upper and the lower confidence limits of the POD and those of the FAR can be computed with the help of many commercially available computer programmes or read from statistical tables [18]. First the calculation of confidence limits for the POD is elaborated.

The number of opportunities to detect a target is marked n, and the number of detections x. The estimated POD is PÔD = x/n. The number of detections is assumed to be binomially distributed with the parameter POD. The two-sided 1 − α confidence limits [20] are

$$\begin{aligned} POD_{lower} &= \frac{x}{x+(n-x+1)F_{1-\alpha/2,\, f_1,\, f_2}} \quad \text{with} \quad f_1 = 2(n-x+1),\ f_2 = 2x \\ POD_{upper} &= \frac{(x+1)F_{1-\alpha/2,\, f_1,\, f_2}}{n-x+(x+1)F_{1-\alpha/2,\, f_1,\, f_2}} \quad \text{with} \quad f_1 = 2(x+1),\ f_2 = 2(n-x) \end{aligned} \qquad \text{(C.1)}$$

The quantities $F_{1-\alpha/2,\, f_1,\, f_2}$ are F-quantiles (also called percentage points) of the F distribution. The difference between the upper and the lower confidence limit is called a confidence interval. Binomial confidence intervals are also called Clopper-Pearson intervals. In the special case when x = 0 the two-sided confidence limits are

$$\begin{aligned} POD_{lower} &= 0 \\ POD_{upper} &= 1-\sqrt[n]{\alpha/2} \end{aligned} \qquad \text{(C.2)}$$

When x = 0, than

$$\begin{aligned} POD_{lower} &= \sqrt[n]{\alpha/2} \\ POD_{upper} &= 1 \end{aligned} \qquad \text{(C.3)}$$

The usual choice is α = 0,05, that is, 95% confidence limits.

The confidence limits for the FAR are described in the following lines. The number of false alarms in a single run follows a Poisson distribution. A variable which is a sum of Poisson distributed variables also follows a Poisson distribution. Consequently, the total number of false alarms y in N repeated scans over an area of size A also follows a Poisson distribution. The estimated false alarm rate is y/(NA). The two-sided confidence limits are [21]

$$
\begin{aligned}
FAR_{lower} &= \frac{1}{2NA}\chi^2{}_{\alpha/2,\,f} \quad \text{with} \quad f = 2y \\
FAR_{upper} &= \frac{1}{2NA}\chi^2{}_{1-\alpha/2,\,f} \quad \text{with} \quad f = 2(y+1)
\end{aligned}
\qquad \text{(C.4)}
$$

where $\chi^2{}_{\alpha/2,\,f}$ and $\chi^2{}_{1-\alpha/2,\,f}$ are called quantiles or probability points of the χ^2-distribution. In the special case when y=0, the confidence limits are

$$
\begin{aligned}
FAR_{lower} &= 0 \\
FAR_{upper} &= \ln(2/\alpha)
\end{aligned}
\qquad \text{(C.5)}
$$

Equations from (C.1) to (C.5) shall be used to calculate the 95% confidence limits for the POD and the FAR.

Before the widespread use of computers some approximative procedures had been developed. However, today their use seems to be hardly justified, since even the most common spreadsheet programmes can deal with the F-distribution and χ^2-distribution, functions necessary for computing the confidence limits of a binomial and a Poisson distributed variable. However, two extreme approximations can be helpful for a quick preliminary assessment of the size of the confidence intervals.

New abbreviations are introduced: p = POD and q = 1 – p. It has been proposed [18] that a normal approximation of a binomial distribution can be used if n > 5 and

$$\frac{\left|\sqrt{\frac{p}{q}} - \sqrt{\frac{q}{p}}\right|}{\sqrt{n}} = \frac{|p-q|}{\sqrt{npq}} < 0{,}3 \qquad \text{(C.6)}$$

This condition means that n is sufficiently large and p=POD is sufficiently far from 1 and 0. For the confidence level $1 - \alpha = 95\%$ and with some additional approximations [20] the following relation holds:

$$POD_{upper/lower} = p \pm 2\sqrt{\frac{pq}{n-1}} \qquad \text{(C.7)}$$

If p is close to 0,5, further approximation is possible:

$$POD_{upper/lower} = p \pm \frac{1}{\sqrt{n}} \qquad \text{(C.8)}$$

When y exceeds 15, the Poisson distribution can be approximated by a normal distribution [21]. Setting the variance of that normal distribution to be equal to the variance of the Poisson distribution, we get approximate 95% confidence limits for the FAR:

$$FAR_{upper\,/\,lower} = F\hat{A}R \pm 2\sqrt{\frac{F\hat{A}R}{NA}} \qquad \text{(C.9)}$$

where FÂR = y/NA is the estimated FAR.

Because of their simplicity, equations (C.7), (C.8) and (C.9) can be helpful in the preparation phase of the experiment. Normally equations from (C.1) to (C.5) shall be used in trial reports, since F-quantiles and χ^2-quantiles can be easily calculated.

Annex D: POD curves

The detection of a target is modelled as a Bernoulli experiment where the binary random variable Y takes its value y = 1 (“detected”) with the probability p and its value y = 0 (“not detected”) with the probability 1−p. The parameter p is specific for each choice of operator-dual sensor-lane combination and it depends on the influence variables characterising that treatment. We cannot relate p linearly with the influence variables, since p is limited to $0 \leq p \leq 1$. Therefore, the parameter p of the Bernoulli distribution is transformed into the parameter η:

$$\eta = \ln \frac{p}{1-p} \qquad \text{(D.1)}$$

This transformation is called logistic (or logit) transformation and the inverse function

$$p = \frac{1}{1+e^{-\eta}} \qquad \text{(D.2)}$$

is called the logistic function. It is a monotonically increasing S-shaped curve starting with $p(-\infty) = 0$ and ending with $p(\infty) = 1$. The parameter η, which is between −1 and +1, is linearly related to the influence variables:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m = \sum_{j=0}^{m} \beta_j x_j \qquad \text{(D.3)}$$

where one of the x_j ’s stands for the depth of the target and the other x_j ’s are indicator variables with values 1 or 0 indicating the presence of a particular level of a qualitative factor. This model is called a generalised linear model [22]. The unknown parameters β_j of the generalised linear model are estimated by maximum likelihood estimation. The result is a curve of POD versus target depth for each combination of other factor levels. The calculation of confidence bounds to POD curves is described in [3].

Bibliography

[1] CWA 14747:2003, CEN Workshop Agreement, Humanitarian Mine Action — Test and Evaluation — Metal Detectors, June 2003. Available at http://www.itep.ws/.

[2] C. Mueller, M. Gaal, M. Scharmach, U. Ewert, A. Lewis, T. Bloodworth, P.-T. Wilrich, and D. Guelle, Reliability Model for Test and Evaluation of Metal Detectors. Final report of the ITEP project 2.1.1.2, Federal Institute for Materials Research and Testing (BAM), Berlin, Germany, Sep 2004. Available at http://www.itep.ws/.

[3] C. Mueller et al, Reliability Model for Test and Evaluation of Metal Detectors — Part 2. Final report of the ITEP Project 2.1.1.8, Federal Institute for Materials Research and Testing (BAM), Berlin, Germany. To be published at http://www.itep.ws/.

[4] M. Gaal, Trial Design for Testing and Evaluation in Humanitarian Mine Clearance, a doctorate thesis submitted to BTU Cottbus in January 2007. To be published at http://www.itep.ws/.

[5] D. M. Guelle, A. M. Lewis, M. A. Pike, A. Carruthers, and S. M. Bowen, Systematic Test and Evaluation of Metal Detectors (STEMD). Interim Report Field Trial Lao, 27th September - 5th November 2004. Technical report, JRC, Joint Research Centre of the European Commission, March 2005. Available at http://www.itep.ws/.

[6] D. M. Guelle, A. M. Lewis, M. A. Pike, and C. Craill. Systematic Test and Evaluation of Metal Detectors (STEMD). Interim Report Field Trial Mozambique, 12th April - 5th May 2005. Technical report, JRC, Joint Research Centre of the European Commission, November 2005. Available at http://www.itep.ws/.

[7] D. M. Guelle, A. M. Lewis, and P. Ripka. Metal Detector Trials — Detector Test Results and Their Interpretation. Technical report, JRC, Joint Research Centre of the European Commission, 2006. Available at http://www.itep.ws/.

[8] A. M. Lewis, T. J. Bloodworth, D. M. Guelle, F. R. Littmann, A. Logreco, and M. A. Pike, Systematic Test & Evaluation of Metal Detectors (STEMD). Interim Report Laboratory Tests Italy. Technical report, JRC, Joint Research Centre of the European Commission, 2006. Available at http://www.itep.ws/.

[9] Geneva International Centre for Humanitarian Demining, Metal Detectors and PPE Catalogue 2007. Available at http://www.gichd.org/.

[10] D. Guelle, M. Gaal, M. Bertovic, C. Mueller, M. Scharmach, M. Pavlovic, South-East Europe Interim Report Field Trial Croatia, March 2007. Available at http://www.itep.ws/.

[11] A. M. Lewis, Test and Evaluation of Dual Sensor Mine Detectors based on a Combination of Metal Detector and Ground Penetrating Radar, technical report, Joint Research Centre of the European Commission, October 2006. Available at http://www.itep.ws/.

[12] R. C. Doheny, S. Burke, R. Cresci, P. Ngan, R. Walls, Handheld Standoff Mine Detection System (HSTAMIDS) field evaluation in Thailand, SPIE, Detection and Remediation Technologies for Mines and Minelike Targets X, Florida, USA, 2005. Available at http://www.itep.ws/.

[13] R. C. Doheny, S. Burke, R. Cresci, P. Ngan, R. Walls, J. Chernoff, Handheld Standoff Mine Detection System (HSTAMIDS) field evaluation in Namibia, SPIE, Detection and Remediation Technologies for Mines and Minelike Targets XI, Florida, USA, 2006. Available at http://www.itep.ws/.

[14] ERA Technology Ltd, MINEHOUND^{TM} Trials 2005-2006, summary report, October 2006. Available at http://www.itep.ws/.

[15] I. M. Dibsdall, Emerging Standards for Testing of Multisensor Mine Detectors, SPIE, Detection and Remediation Technologies for Mines and Minelike Targets X, Florida, USA, 2005. Available at http://www.itep.ws/.

[16] J. B. Rhebergen, J. M. Ralston, Test and Evaluation Protocols for GPR-based Mine-detection Systems; a Proposal, SPIE, Detection and Remediation Technologies for Mines and Minelike Targets X, Florida, USA, 2005. Available at http://www.itep.ws/.

[17] D. C. Montgomery, Design and Analysis of Experiments, John Wiley & Sons, third edition, 1991.

[18] G. E. P. Box, W. G. Hunter, J. S. Hunter, Statistics for Experimenters, John Wiley & Sons, 1978.

[19] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 3 April 2007.

[20] K. M. Simonson, Statistical Considerations in Designing Tests of Mine Detection Systems: I — Measures Related to the Probability of Detection, technical report, Sandia National Laboratories, August 1998. Available at http://www.uxocoe.brtrc.com/techlibrary/TechRpts/misc3.asp.

[21] K. M. Simonson, Statistical Considerations in Designing Tests of Mine Detection Systems: II — Measures Related to the False Alarm Rate, technical report, Sandia National Laboratories, August 1998. Available at http://www.uxocoe.brtrc.com/techlibrary/TechRpts/misc4.asp.

[22] StatSoft Inc., Electronic Statistics Textbook, Tulsa, OK, USA, 2006, http://www.statsoft.com/textbook/stathome.html.