HECTOS

Harmonized Evaluation, Certification and Testing Of Security products

# D5.4
# Testing and Evaluation results for the Weapons and Explosives detection case studies

## *EXECUTIVE PUBLISHABLE SUMMARY*

A report prepared by:
**FhG ICT**
**TNO**
**Iconal Technologies**

Date: 16.05.2017
Project No: 606861
FOI Designation No: FOI-2015-1958
Dissemination Level: PU(Summary)
Total No of Pages: 9

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 606861.

# D5.4
# Testing and Evaluation results for the Weapons and Explosives detection case studies

| | |
|---|---|
| Version: | 2.0 – Executive Publishable Summary |
| FOI designation no: | FOI-2015-1958 |
| Responsible: | FhG ICT |
| Author(s): | Christian Ulrich, FhG ICT<br>Marta Jezierska-Switala, Martijn Koolloos, Erik Kroon, TNO<br>Mike Kemp, Sam Pollock, Iconal Technologies |
| Number of pages: | 17 |
| Dissemination level: | *PU – valid for the Executive summary* |
| Start date of project: | Sep, 2014 |
| Duration: | 41 months |

## Summary

HECTOS is an EU FP7 security research project exploring the issue that there are very few evaluation and certification procedures for physical security products that are mutually recognized by EU Member States. HECTOS intends to identify mechanisms to evaluate the performance of security products, also taking into account aspects such as interoperability requirements, regulatory requirements, ethical requirements, and privacy requirements. The project will propose elements of a roadmap for the development of new harmonized product certification schemes.

To analyse, develop, enhance, and experimentally investigate evaluation and certification schemes, HECTOS conducts case studies in two priority areas: Biometrics and detection of weapons and explosives. This report is an executive publishable summary of the results from the weapons and explosives detection systems case studies and is divided in three independent parts:

**Part I:** Explosives Trace Detection Products

This part focuses on repeatability of test results for explosives trace detection equipment, with the objective to identify important aspects that ensure inter-lab and intra-lab repeatability. This was investigated by developing a basic test method composed of five test blocks which was implemented independently at two laboratories at ICT and TNO. Particular studies have been carried out on sample preparation, sampling from surface and testing of volatile explosives.

**Part II:** People Screening Portals

This part focuses on the following topics related to screening of persons

1. Determination of the Receiver Operator Characteristic (ROC) curve, with the objective to investigate application-driven evaluation

2. Verification of technology independency, with the objective to identify elements of the test method that enhance or impede technology independency

The ROC curve determination revealed that test person-based testing gives a more realistic assessment of the Walk-Through Metal Detector (WTMD) than the tests without test persons but at the expense of accuracy and repeatability. Furthermore, it is much more convincing to consider the WTMD as a simple binary detection system rather than using performance assessment with zone indication. The ROC-curve based evaluation seems to be especially useful when operated in partially divested mode. For non-divested mode full detection is obtained, even at the lowest sensitivity and with and without threat items. For full divested operation, the False alarm rate is zero and no ROC-curve is obtained.

**Part III:** Testing of Early Stage Technology

This part describes the development of approaches to the testing of low-TRL prototypes of physical security products. The approach was supported by two case studies focusing on detection technologies (Raman-based particle trace detection and active millimetre wave people screening) with an additional discussion on the application of the approach to other types of products.

# Part I: Explosives Trace Detection Products

D5.4 Part I summarises the outcome of various tests that have been conducted to investigate the critical elements for harmonised testing of explosive trace detection (ETD) equipment identified in HECTOS D5.2.

The main elements that have been investigated are

- Sample preparation
- Sampling from surface
- Testing volatile explosives
- Repeatability of testing

The latter included the inter-laboratory repeatability investigated by developing a basic test method composed of five test blocks which was implemented independently at two laboratories at ICT and TNO.

Other elements relevant for the harmonised evaluation of ETD systems have also been taken into account.

## *T & E of High TRL Explosive Trace Detection equipment – the approach*

Explosives trace detectors (ETD) are security detection equipment indicating the presence of explosives by detecting trace amounts (nanogrammes to microgrammes) of explosives, sampled in form of particles or as a vapour. Trace detectors only indirectly detect the explosives and rather give the basis to assume that the scrutinised object or a person came into contact with explosives, than provide the evidence of a direct explosive threat. In order to provide well documented approach for testing and evaluation in HECTOS, the focus has been put on particle sampling devices. It has been chosen to use an instrument based on a well-known and widely present Ion Mobility Spectroscopy (IMS) technique. Various particle IMS trace detectors, are present on airports, checkpoints or for example prisons, where they are used as a secondary screening tool or in the alarm recognition phase.

In concepts of operations (CONOPs) of the most trace detectors particles of explosives are sampled by wiping the contaminated surfaces with special and dedicated swabs. Typically the swabs are then heated inside the detector to facilitate the evaporation process. The evaporated molecules are ionised and analysed using a detection device. In case of IMS sensors, the molecules can additionally be identified by comparison with the library present. For other instrument present on the market the identification is not always possible.

The work of preparing and performing the case study Explosives Trace Detection (ETD) focused mainly on experimental evaluation of elements described in the Test Methodology of ETD equipment that have been identified as essential for a harmonised approach. A harmonised evaluation procedure is of fundamental importance in order obtain repeatability over time and among test laboratories leading to a well-accepted scheme in the EU.

In D5.2 the necessary contents of test methods for ETD systems have been derived on a high level. Transferring these contents in a detailed description that can be used in real testing has been done for the work described in the present report.

The following approach has been chosen for the current case study:

- Identify the partner labs for the repeatability testing.

  - Within a lab separate the specialists working on a dedicated Test Method from the practitioners that should execute the test in the later study.
  - Determine the stepwise workload of the test case, introducing the goals for each step.
- Identify the elements having impact on the harmonised testing of the trace detection equipment based on accessible investigations, literature studies and own experience;
  - Identify challenging substances and the way they should be tested.
- Identify the trace detection equipment to take part in the trial.
- Perform preliminary experiments for major impact elements underlying their applicability in the Test Method. Consider materials for both: alarm and false alarm.
- Based on the results of the experiments and the literature investigation decide which experiments will be a part of the repeatability test. Having in mind properties of the explosive samples decide which particular explosives should be tested. Decide on the number of repetition and combination of tests.
- Prepare a detailed Test Method, introducing all necessary elements and the way of performing measurements, with almost no room for individual “guessing” but relying on good laboratory practice and trace detection experience of the practitioners.
  - The test Method should be as generic as possible.
  - Agree on identical substances, materials and accessories to be used.
- Execute the Test Method in the first laboratory.
  - During the testing collect two kinds of the remarks: on the clarity and transparency of the Test Method and on the individual test outcomes and behaviour of the instrument.
- Move the equipment to the second laboratory without revealing the outcomes and execute the trials.
- Combine the results and analyse the outcomes.
  - Interview the practitioners on their experiences with the test and the Test Method and equipment manufacturer if necessary.
- Draw the conclusions for the test case.
- Interpolate the test case conclusions in comprehensive HECTOS approach (to be performed in the later stage).

## *Elements with major impact on harmonised testing of ETD*

The following elements of testing have been identified to have major impact on the results of testing and were investigated in detail in this case study.

### Sample preparation

Trace amounts of explosives in the range of tenth to hundreds of Nanogrammes cannot easily be deposited on substrates or sample swabs as solid substances, as such little amounts cannot be weighed and transferred. Usually this problem is solved by using diluted solutions of the explosives in suitable solvents, as liquids can be deposited in a reproducible way using pipettes or syringes. After transferring a solution to the surface of interest or sampling swab, the solvent evaporates, while the less volatile explosive remains on the substrate surface.

Owing to various properties of explosive materials, the direct liquid transfer to surfaces has some disadvantages:

- Volatile explosives for example EGDN, NG or TATP evaporate within the same timescale as the organic solvents they are dissolved in. So the amount of these substances decreases very fast and unpredictably after deposition and the actual amount left on the surface cannot be guaranteed. Transferring of volatile substances from solutions to test surfaces with subsequent sampling from those surfaces is therefore not suitable. Testing of volatile substances is discussed below.
- In porous substrates the transferred liquids permeate in the material by capillary forces leading to residues of explosives spread through the solid material and its pores instead of keeping them on the surface. Therefore the explosive material may partially not be collected by sampling the surface.
- Some surfaces are not resistant to organic solvents needed for the explosive solutions; the explosive can therefore partially be covered by solvated surface material or included in the substrate dissolved by the solvent.

These problems are solved for low volatile substances in Patent N0.2 US 6,470,730 B1 by using a second, inert Polytetrafluoroethylene (PTFE)-surface, where the explosive solution is deposited and can be transferred from this surface after drying. This procedure is called the Bytac-transfer as Bytac® foil is used as transfer media which consists of a thin PTFE-layer on self-adhesive support backing made from Polyvinylchloride (PVC).

But there are restrictions lowering the advantage of this dry transfer:

- The main amount of solvent evaporates very soon leaving a visible dry residue on the PTFE surface. But microscopic studies at ICT showed that this residue consists of a saturated solution from which the remaining solvent needs a very long time to evaporate. Depending on solvent and explosive the crystallisation of the residue can take several days. This means that after drying times of several minutes to some hours there might be still no dry particles and still a liquid is transferred.
- The residue on the transfer-surface has to be transferred to the surface of interest. The efficiency of this transfer step is unknown and depends strongly on morphology and kind of explosive and of both surfaces.

The force applied for transfer also influences the efficiency. In preliminary tests at ICT the Parared dye, which is much better visible under the microscope than usual explosives, was transferred from Bytac to different surfaces with varying forces on a balance. These experiments showed that the transfer efficiency increased with increasing load and was higher with rougher surfaces. Increasing the load over 30 N led to damages in the PTFE-layer of the Bytac foil.

Similar experiments have been performed at TNO using RDX solution. Those experiments proved that the force of 20 N should be applied to transfer the crystals from the Bytac to the soft surface as cardboard. By using lower force (10 N), the majority of crystals was still observed on the surface of the Bytac. Using higher force was found to be unpractical and difficult to practice.

### Sampling from surfaces

Trace contaminations of explosives are the result of handling these explosives leading to particle contamination of hands, clothing or other objects. These particles can unwillingly be transferred to other belongings touched with the contaminated hands or clothes. Surfaces which are frequently touched like for example mobile touch screens, computer keyboards, wallets, handles of suitcases and bags, zippers, pockets and spectacle frames are most likely the surfaces of interest to search for such contaminations in order to find the “bad guys” planning an explosive attack.

Therefore the goal for ETD-systems is collecting explosive particles from various surfaces and subsequently detecting them.

Testing ETD obviously needs samples of various surfaces with known trace amounts of various explosives.

Besides the difficulty to get the trace on the surface of interest, the second difficulty is to effectively sample the trace from the surface. This sampling efficiency depends on the chemical nature of trace substance and surface as well as the morphology of trace substance and surface and not to forget the roughness and morphology of the swab used for collecting. As manufacturers usually deliver ETD with dedicated swab materials, this parameter was not varied in the course of this exercise. The influence of the test surface and the interaction between surface and explosive traces was examined in various combinations.

### Testing volatile explosives

The amount of volatile substances transferred to substrates decreases very fast due to evaporative loss. The situation is even worse when small amounts of substances are transferred in a solvent as the evaporation of the solvent and of the volatile substance occur in parallel.

Therefore, the desired amounts of substances which are in the same region of volatility (EGDN, TATP, NGL) cannot be guaranteed for testing ETD-systems. The only possibility to test those substances seems to analyse them shortly after application. The major disadvantage of this is that the detector has to detect trace amounts of the substance while much bigger amounts of the solvents are still present as 1 µl solvent corresponds (according to its density) to almost 1 mg while the trace amount of explosive is in the scale of Microgrammes to tenth of Nanogrammes. This huge amount of substance can not only lead to an overload of the trace detectors but often leads to false alarms of ETD. (For example in IMS-systems the evaporation of 1 µl solvent in the desorber changes the temperature in the drift tube what may lead to false alarms).

Other detection technologies also struggle with those big amounts of solvents.

In consequence volatile explosives can only be tested in solvents which do not cause false alarms and the optimal drying times for each combination of substance and solvent has to be examined. The balance of remaining enough explosive for detection with a minimum of interference through the solvent has to be found for each single system as the evaporation speed depends not only on temperature and moisture, but also on the morphology and surface of the swabbing material used for introducing the samples into the detector.

In the current exercises two volatile explosives were chosen to be tested in the interlaboratory test, where both labs tested exactly the same system with the same swabs using the same solvent and the same concentration of the solutions.

Despite the problems with solvents described above this is only an issue for laboratory testing, but not for operational conditions. When seeking traces of volatile explosives in a real environment as an indication of a hidden threat, solvents are usually not present, as the explosives are used as bare materials.

### Repeatability of testing - interlaboratory

An established method for the evaluation of test methods are so called “round robins” where different laboratories perform tests using the same test methodology (e.g. with exactly the same

materials and procedures etc.) and the results are used to evaluate the accuracy and repeatability of this test methodology. These kinds of tests are furthermore used in the quality control for example in accreditation following ISO 17025 to show that a laboratory is proficient to perform tests according to evaluated test methods with the required accuracy.
The WP5 partners agreed that a reduced round robin with two participating test laboratories will be a useful support estimating the influence of human factors on the repeatability of test results. This includes the variability of results performing the same tests at different laboratories as well as variability on the interpretation of written test methods by different users.

A test method (TM) for the interlaboratory comparison was created by the trace detection specialists at ICT and TNO. Many factors which could influence repeatable testing were considered and implemented in the TM and all steps were described in detail. The use of certified Standard Reference Materials (SRM) - where available - was appointed, the surface materials were purchased by one of the labs and distributed to both to guarantee that exactly the same materials are used. To ensure the neutral evaluation of this test method it was decided, that the personnel conducting the tests was not involved in the preparation of the test method.
The TM was composed of five test blocks:

1. Direct deposition of explosive solutions on sample trap followed by thermal desorption and analysis in the ETD. This block includes the determination of optimum evaporation times for volatile explosives. This part is performed with six different explosives.
2. Surface sampling: The explosives have to be transferred to substrates and sampled from these surfaces. This part is performed with five different explosives on three different substrates. This block includes the description of the dry transfer procedure.
3. Testing of benign items by direct deposition. The results are used to calculate a laboratory false alarm rate when the substances are directly introduced to the ETD. This part is performed with 4 benign items purchased as standard reference materials (SRM).
4. Testing benign items on surfaces. This part enlarges part 3 with the possible interference of the substances and the substrates and includes the dry transfer and sampling from surface. The same four benign items are tested on three surfaces.
5. Suppression test: Testing the combination of explosive and benign item being present on the same sample trap as the most complex test showing the capability of detection in the presence of background challenging materials (BCM).

The aim in preparing the TM was to reveal factors with main influence on the repeatability of trace tests. For this purpose the selection of kind and amount of explosives, substrates and benign items has been made with the view on consistent usability and is not reflecting any legislative specification or sensitive information. Furthermore, the extent of the single test blocks was chosen to be big enough to prove the concept, without claiming to be complete.

The test method was implemented independently in two laboratories: in July and August 2016 at ICT (Germany), and in September 2016 at TNO (The Netherlands). The tests were performed with a handheld IMS based trace detector, the Bruker "RoadRunner", kindly provided by Bruker Daltonik GmbH, Leipzig. Exactly the same instrument (transported between both labs by its manufacturer) was tested in both locations; the same materials and requisites have been used. The results acquired during the first test at ICT have not been shared with the second laboratory performing the test until the whole exercise was finished.
The comparison of the test results showed partially significant deviations. The reasons for this could not in all cases be identified yet, and will be further explored in WE5.5.

One reason for the deviations was surely that the testing has been performed with really small quantities. The system was challenged to work in the region of its detection limit leading to some detection rates considerably below 100%. The differences in the test results would have been much smaller when either the amount had been chosen higher (DR in both labs near to 100%) or lower (DR in both labs near to 0%). Furthermore, the application of the TM by users not involved in their preparation revealed some formulations that have to be made more precise when writing a TM. Preparing a real TM for harmonised testing therefore requires more iterative loops of formulating and implementing the procedure in pilot tests.

**Repeatability of testing - intralaboratory**

Intralaboratory repeatability refers to the consistency of test results within one lab over time. It also includes differences found due to different skills of the operators taking part in the study.
Many of the tests from the TM for interlaboratory study have also been performed by different members of the ICT laboratory team. The direct deposition tests showed no significant differences between three operators.

# Part II: People screening portals

The objective D5.4 Part II is to elaborate a detailed test methodology (TM) for Weapon and Explosives (W&E) detection devices (portals) for people screening. This TM is then used to carry out different tests for investigating elements of the harmonized certification scheme. It concerns topics identified in D3.2 as specifically important to be included in such a scheme:

- Determination of the ROC-curve, with the objective to investigate the application-driven evaluation;
- Repeatability, with the objective to identify important aspects that ensure intra-lab repeatability;
- Verification of technology independence, with the objective to identify elements of the test method that enhance or impede technology independence.

For the ROC-curve determination (first topic) and the intra-lab repeatability (second topic), performance evaluation methodologies have been specified, based on the general test method developed in D5.3. For both test methods, real test persons were used. The technology independence verification (third topic) is assessed by means of a theoretical study.

## *Determination of the ROC Curve*

The security performance of a detector is commonly described by means of the probability of detection (Pd) and the probability of a false alarm (Pfa). These probabilities are not exactly predictable by testing. Testing results consist of values for the detection rate (DR) and the false alarm rate (FAR). DR and FAR are the estimators of Pd and Pfa. The DR is determined from the number of alarms when the portal is tested with threat runs, while the FAR is determined from the number of alarms when the portals is tested with benign runs. The accuracy of the test results and hence of the prediction of detection probabilities generally increases with realistic ways of testing and with test sample size.

For people screening portals the FAR depends strongly on the level of divesting of a person combined with the sensitivity settings of the portal. During testing the level of divesting is optimized, however during normal operation the level of divesting depends on the preparation of people to be scanned (e.g., removing metal items as they approach the detector). This may results in large discrepancies between laboratory assessed FAR and operational FAR. Since the operational FAR has a direct effect on the throughput of the overall process (portal + alarm resolution), it forms an important user requirement, which can have an important effect on the test method.

Normally, there is a trade-off between DR (benefits) and FAR (costs), which depends on the detector settings (sensitivity). The sensitivity settings of the machine is adjustable and for different settings, different combinations of DR and FAR are obtained. All points together form the so-called receiver operating characteristic (ROC) curve of a detector. The closer a result is to the upper left corner of the ROC space (i.e. high detection rate versus low false alarm rate), the better the machine's detection performance in terms of sensitivity and specificity. Therefore, the distance from the random guess line (diagonally running from 0% detection / 0% false alarms to 100% detection / 100% false alarms), to the upper left part of a ROC curve (the ideal test), also indicated as d' (d prime), is an illustrative indicator of the performance of a screening device. Normally, a detector shows different ROC curves for different object sizes. This may have two implications:

1. For the same setting, the detection rate decreases with decreasing object size;
2. In order to maintain the same detection rate for decreasing object size, the settings have to be changed and the false alarm rate will increase.

The performance requirements of a people screening portal may differ significantly, depending on the application. For some applications it is very important to detect all threat objects, including the small ones and the fact that this comes at the cost of a high FAR is less important. For other applications only large threats may be of special interest and it may be more important to have a very high throughput and hence a low FAR. Also, the operationally feasible level of divestment plays a role. For certain application scenarios like for example Educational facilities there may not be any divestment, while for others (Courthouses or Nuclear facilities) limited or full divestment may apply.

The test described in this report explored performance evaluation by determining the ROC curves for several object size classes, using a WTMD with alarm zone indication. Moreover, a person-based test protocol is used instead of the mechanical frame-based tests. This is done with the aim to use the general base evaluation test protocol for all people screening portals from D5.3.

The table below gives an overview of the controlled variables for the WTMD ROC-curve testing.

| Machine sensitivity | Level of Divestment | Test persons | | Threat item | | |
|---|---|---|---|---|---|---|
| | | Gender | BMI | Threat ID | Size class | Location |

The orientation of items has not been implemented as an independent variable, but rather as a random variable, depending on the test person, the threat morphology and the location.

From this test campaign, the following conclusions were drawn:

- For non-divested runs (i.e. including purse and small back pack) with and without threat items, the WTMD used in the case study gave maximum alarm, even at the lowest sensitivity. Hence, this modus operandi seems not practicable, but should be scope-tested for each type of WTMD.
- The spatial discrimination of the WTMD used in the case study is not good enough to support performance assessment with zone indication (CTA, TFA, IFA). It is much more convincing just to consider the WTMD as a simple binary detection system (alarm / no alarm). The zone information could then be used to guide a complete pat-down, rather than an indication for targeted pat-down.
- All persons within group “partial divested” should carry exactly the same items. It is recommended to define several levels of “partial divested” with well-defined differences of object sizes.
- A careful selection of the innocuous item set is very important since it determines the FAR, i.e. horizontal axis in ROC curve.
- There needs to be a wide variety of test persons (gender, BMI) and all test persons must do the same number of runs to ensure that the specific performance of individual test persons (+specific innocuous objects) does not bias the results of the test. As for the SSc

(see chapter 3), it is thought that 16 different test persons will be sufficient, but this requires further investigations.

- Very small threat items (e.g. T06) may not be detected[1] in full divestment mode, but in partial divestment modes this kind of threat items may be found because of a pat down after a false alarm indication (but this depends clearly on the screener's performance).
- The false alarm rate for fully divested persons is 0%. However, small innocuous items that contain metal, such as glasses or the metal in a bra strap may generate an innocuous false alarm at high machine sensitivities.
- The situation that a single item (either threat or innocuous) does not trigger an alarm when passing through the portal, but does trigger an alarm when another item (either threat or innocuous) is carried, occurred many times. Therefore:
    - Testing with only single threat items may be suitable for comparison of detection performance of portals, but does not give a realistic impression of the actual detection performance in the field.
    - Partial divestment may result in higher detection rates compared to fully divested screening.

### *Intralaboratory repeatability*

The objective of the intra-lab repeatability test is to identify important elements that ensure intra-lab repeatability. For people screening portals it is known from experience that the human factor in the test, i.e. the test person, is the main cause of variance. Since the Security Scanner is a commercial product for which a lot of evaluation experience is available within HECTOS and since from a perspective of technology independence testing (see chapter 4) there is a wish to apply the general test method from D5.3 on different types of portals, it was decided to investigate the intra-lab repeatability by using a Security Scanner. Two test series were carried out with some time in between the tests and by using different test managers. For both test series the test managers were responsible to carry out the test according to the same test protocol. To reflect screening conditions in practice (with parameters independent of the test protocol), the tests may differ in test persons, the exact way of carrying out the test and alarm indication interpretation. The test persons were selected by the test manager. Informed Consent Forms were provided to each test person which he/she had to read and sign.

Security scanners are capable of detecting small anomalies on the body (e.g. a watch) or underneath clothing (both threat items and benign items), which are indicated on a screen by detailed threat location information, and do not show interference of different (threat) items on different locations on the body. Therefore, for detection performance tests without alarm resolution by a human screener (so where a correct alarm is deduced only from the information on the screen), it has no added value to apply partially divested runs: an alarm indication collocated within the zone where the actual threat item is worn is counted as a correct alarm and all other alarms are disregarded, irrespective of whether these are caused by an innocuous item or whether these are true false alarms. Consequently, only fully divested runs were applied. If the test person used glasses he/she could wear those glasses during the test. Male test persons were allowed to keep the belt during the test, but only when no consistent alarm was generated by the belt.

[1] For present tests with the specific WTMD, but for other machines such items may be detected.

The number of runs were such that a 95% Confidence Interval (95%CI) of 5% was obtained on parameter granularity level (i.e. per gender/BMI combination or per threat item). Furthermore, in order to be able to distinguish differences between the two test series, the threat item / location combinations were selected such that the Detection Rate (DR) and False Alarm Rate (FAR) were reasonably far from 0% or 100%.

The table below gives an overview of the controlled variables for the SSc intra-lab repeatability testing. Two test series were done with 8 test persons. Possible differences in detection performance of the SSc between a left turn of the scanning beam and a right turn (whether this effect actually occurs was not investigated) were averaged by performing 4 runs for each unique parameter setting. In total 600 FA runs and 1280 threat runs were done per test series.

| Test persons | | Threat item | | |
|---|---|---|---|---|
| Gender | BMI | Threat ID | Size class | Location |

The effect of different clothing was not included in this research. Threat items were concealed by clothing and attached to body by using the clothing itself (e.g. a small weapon in sock on ankle), or elastic band (if using clothing is impracticable). The test manager determined together with the test person the exact location and orientation and the way the threat item was attached to the body.

Statistically significant differences were observed between the two test series, both for the false alarm runs and for the threat runs. However, at first sight the differences were small and the same trends were observed for both test series. At the granularity level of gender / BMI combination the differences between the two test series was only for one combination statistically significant. These results indicate that the repeatability of the test as performed during this study was reasonably good, where “good” would mean that there are no significant differences and “bad” would mean that almost all differences were statistically significant and the observed trends did not agree.

During the tests it was observed that the test managers gave slightly different instructions with respect to the way the threat items were attached to the body and the test persons were more or less free to choose the most convenient way. It was also observed that the clothing of the test persons varied considerably: two test persons in series A were in uniform, where all other test persons were in casual, but with the variance that one can expect to observe at an operational checkpoint.

Given these considerations, the observed differences can only be caused by differences in:

- Test persons for same gender / BMI combinations
- Clothing (concealment)
- Position and orientation of threat items
- Attachment to the body of threat items: strapped, taped or fixed by clothing (e.g. by tight shirt, jeans or bra strap)

Since the number of runs were statistically justified, it is thought that the repeatability can be increased when one or more of the following measures are included in the test protocol:

- Same number of runs, same number of threat items but an increased number of test persons (>16).
- Prescribed type of garment (casual / business)
- Prescribe the exact location and orientation of the threat item.
- Prescribe in a detailed way how to attach the threat item to the body.

The effectiveness for the repeatability on all these measures require further investigations in order to develop a good TM.

### *Technology Independence*

At first sight, it might be thought that a test method for a system to detect forbidden items could be written in way that is independent of the particular detection technology. Guns, knives and explosives of different sizes can be concealed on people in different locations on the body. The persons then pass through the detection system and the detection rate is calculated.  Additional factors such as the ability to discriminate between threat and benign items, the effect of walking speed etc., can easily be introduced.

As observed during the tests described above, the test method for a SSc differs at some aspects from that for a WTMD. Although the base of the TM with test persons as described in D5.3 can be used for both technologies, the detailed test protocol may show some important differences due to the different technologies they are based on. As a further example, this reports gives a theoretical evaluation of the applicability of the test method used for the WTMD to a different type of metal detector which has a different ConOps and, more importantly, uses a different detection technology.  This is a so-called Walk By Metal Detector (WBMD) using passive magnetometry as its principle of detection. This question is of particular interest since although the name suggests only a slight difference between the two detectors (walk by or walk through), the technology on which the detectors are based differs considerably.

The principal conclusion that was drawn is that care needs to be taken to ensure that the assumptions made in the development of a test method are valid for all the different types of equipment that may be tested.  In particular, they need to be examined when a new technology is introduced which uses different physical or chemical principles and corresponding ‘signatures’. All test methods are designed with some assumptions about the parameters of threat objects, test subjects, and environmental conditions that need to be taken into account. If a detector is introduced which uses a different detection signature, these assumptions (which are often implicit) need to be identified and challenged. In the case of the two different types of ‘metal detector’ technology, induction metal detection as used in WTMD and passive magnetometry, as used in some recent ‘WBMD’, the assumption that two threat objects of exactly the same type will have the same signature is not always true.  If this is not taken into account, very erroneous test results will be obtained.

### *General Conclusions*

The main conclusions from the evaluations are:

- Test person-based testing gives a more realistic assessment of the WTMD than the tests without test persons. However, due to a number of deliberately uncontrolled parameters, this more realistic results show a high variance and come hence at the expense of accuracy and repeatability;

- Mixed threat item / innocuous item testing gives a more realistic impression of the actual behavior in the field because of the interference that can occur between two or more items. An important consequence of this interference is that full divested screening might result in lower detection rates;
- The ROC-curve approach enables type certification based on its intended application, i.e. divestment level and sensitivity setting;
- The ROC-curve based evaluation seems to be especially useful when operated in partially divested mode. For non-divested mode full detection is obtained, even at the lowest sensitivity and with and without threat items. For full divested operation, the False alarm rate is zero and no ROC-curve is obtained;
- Scoping tests should be used to determine the regions where the detection system has sensible performance, so that testing efforts can be focused on those and to avoid wasting time on regions where the performance is virtually zero or virtually perfect;
- The intra-lab repeatability of the security scanner test was reasonably good but with some statistically significant differences. Based on observations made during the test, the repeatability is thought to increase if a larger number of test persons are used and when the exact threat item location, orientation and way of attachment to the body is prescribed in the test protocol.
- A one-test protocol-fits-all for people screening detection systems is not realistic. It must be ensured that the assumptions made in the development of a test method are valid for all the different types of equipment that may be tested. The operating principles of the device under test have to be examined to determine whether the proposed test method is suitable or needs adjustment to fit those operating principles.

## Part III: Testing of Early Stage Technology

The objective of this report was to describe the development of approaches to the testing of low-TRL prototypes in the physical security domain, aiming to deliver the following anticipated benefits:

- To enable researchers and technology developers to assess the maturity of their work and the potential of the technology;
- To guide researchers by helping them understand the real world environment their technology will need to operate in and identify the areas where further development work is needed;
- To enable technology developers to communicate effectively with their peers and with their sponsors or funding agencies about the progress and potential of their work

Early in the process it was decided to divide physical security products into two broad categories; technologies which ‘Detect’ and those which ‘Delay’ an adversary. It was identified as necessary to take a different approach towards testing in each case. The majority of the effort in this exercise was focused on developing early stage testing approaches for the ‘Detect’ technologies. This was done using case studies of two different types of explosives and weapons detection technology.

Using background research into other existing early stage testing concepts in both security non-security markets, as well as prior experience of working with early stage security technology, an initial high level approach to the generic low-TRL test methodology was developed. This was tested and refined through the development of two technology specific case studies. The first case was test and evaluation of optical Raman trace explosive detection techniques (authored by FOI) and second case was test and evaluation of millimetre wave techniques used for detecting concealed explosives and weapons (authored by Iconal).

The draft Raman test method was used to test actual hardware prototypes under development at FOI to explore which aspects worked well and identify areas for improvement. The draft millimetre wave test method was reviewed by a number of independent subject matter experts comprising developers, system integrators and end-users.  For both test methods, feedback was very positive.  As well as providing the basis for the methodology development, it is felt that both test methods could be used by the R&D community to support a number of ongoing development projects.

Using the experience and insight gained in the case study process, the generic methodology has been refined and developed to the extent that it now stands alone as, hopefully, an accessible guide for users to understand the principles and apply this approach for a new detection technology development.

It was found through the process of reviewing approaches in other fields that although there were no directly applicable methods, there were lessons that could be learnt, and good generic principles that could be extracted from other early stage testing examples.

The process by which the test methodology was generated was found to work well. By building on the high level initial approach to develop the two test cases, then challenging them through lab testing or peer review, the findings were found to be valuable in refining the generic low-TRL test methodology.

The defining principles were generated at the start of the process as high level goals for the activity, and refined throughout by a process of retrospective review. It is considered that the methodology produced in this report meets these principles, and thus hopefully will act as a useful tool to support and guide the development of new technologies.