NATO STANDARD

AEP-4843

CONSIDERATIONS FOR THE TESTING OF MILITARY SEARCH EQUIPMENT

Edition A, version 1

DECEMBER 2021

NORTH ATLANTIC TREATY ORGANIZATION

ALLIED ENGINEERING PUBLICATION

Published by the
NATO STANDARDIZATION OFFICE (NSO)

INTENTIONALLY BLANK

NORTH ATLANTIC TREATY ORGANIZATION (NATO)

NATO STANDARDIZATION OFFICE (NSO)

NATO LETTER OF PROMULGATION

20 December 2021

1. The enclosed Allied Engineering Publication AEP-4843, Edition A, version 1, CONSIDERATIONS FOR THE TESTING OF MILITARY SEARCH EQUIPMENT, which has been approved by the nations in the NATO ARMY ARMAMENTS GROUP, is promulgated herewith. The recommendation of nations to use this publication is recorded in STANREC 4843.

2. AEP-4843, Edition A, version 1, is effective upon receipt.

3. This NATO standardization document is issued by NATO. In case of reproduction, NATO is to be acknowledged. NATO does not charge any fee for its standardization documents at any stage, which are not intended to be sold. They can be retrieved from the NATO Standardization Document Database ((https://nso.nato.int/nso/) or through your national standardization authorities.

4. This publication shall be handled in accordance with C-M(2002)60.

Dimitrios SIGOULAKIS
Major General, GRC (A)
Director, NATO Standardization Office

INTENTIONALLY BLANK

RESERVED FOR NATIONAL LETTER OF PROMULGATION

INTENTIONALLY BLANK

RECORD OF RESERVATIONS

CHAPTER	RECORD OF RESERVATION BY NATIONS
Note: The reservations listed on this page include only those that were recorded at time of promulgation and may not be complete. Refer to the NATO Standardization Document Database for the complete list of existing reservations.	

INTENTIONALLY BLANK

RECORD OF SPECIFIC RESERVATIONS

[nation]	[detail of reservation]
Note: The reservations listed on this page include only those that were recorded at time of promulgation and may not be complete. Refer to the NATO Standardization Document Database for the complete list of existing reservations.	

INTENTIONALLY BLANK

TABLE OF CONTENTS

REFERENCES

MC 0560/2 Military Committee Policy for Military Engineering

AJP-3.12 Allied Joint Doctrine for Military Engineering

ATP-3.12.1 Allied Tactical Doctrine for Military Engineering

ATP-3.12.2 Allied Tactical Doctrine for Military Search

STANAG 4370 Environmental Testing

TNO 2018 R10157 Bijlage B Metal Detection Benchmark Test Protocol Manual. (Can be requested by e-mail from benchmarkprotocolMS@tno.nl).

CEN Workshop Agreement (CWA) 14747, "Humanitarian Mine Action Test and Evaluation - Part I: Metal Detectors," European Committee for Standardization (CEN), 2003.

GLOSSARY

1. The following definitions are repeated from a recent international effort to standardize the evaluation of metal detection in humanitarian demining.[1] Where appropriate, some definitions have been adapted and generalized for use with non-metal detection systems, while others have been introduced by the authors for completeness.

2. Note, these definitions maybe somewhat at odds with what a military reader may be familiar with, but they are consistent with the intent of a technical evaluation.

3. Fundamentally, a detection in this context is when an object capable of being detected — be it by design, or material construction — is detected. That is, a detection is defined by the *physical operating principles* of the detection technology itself, not by the *intent* of the detector's use. For example, as defined here, uninteresting battlefield metallic clutter would be detected by a military metal detection sensor, although the principal operational intent of that system would be to locate explosive hazards.

4. Further, in a military context, such battlefield metallic clutter would commonly be referred to as a false alarm. However, according to the definition below, a false alarm would only arise when a sensor, operated as intended, sensed an anomaly that would otherwise not be expected to be detected by that technology. If the operator subsequently decided the false alarm constituted a detection, it would be referred to as a false detection.

5. These nuances in definition will be important to bear in mind if and when the results of any subsequent verification tests are relayed to military stakeholders.

Alarm indication A signal to warn of the sensing of an object; the indication is often visual, auditory, and/or vibration. A positive alarm indication is repeatable under the same conditions and is not intermittent. (Adapted from CWA 14747[1])

Alarm indicator The device used to generate the alarm indication; the indication can be visual, auditory, and/or vibration. (Adapted from CWA 14747[1])

Alarming object Any object to which the detector is specifically designed to generate an alarm indication. There are three kinds of alarming objects, which may not be differentiable to the detector: target objects, non-target objects (or clutter), and anomalies.

Anomaly Objects/soil/environmental disturbance or intrinsic sensor noise that causes an alarm that would otherwise not be expected to generate an alarm indication based upon the physical operating characteristics of the detection technology being used.

Blind test A test in which the detector operator does not know details of the location, orientation, depth, height, or nature of the target(s) being sought[1].

[1] CEN Workshop Agreement (CWA) 14747, "Humanitarian Mine Action Test and Evaluation – Part I: Metal Detectors," European Committee for Standardization (CEN), 2003

Clutter See non-target object.

Detector A device or instrument that emits an alarm indication based on the output of an internal sensor, or sensors. A detector is the practical embodiment of a sensing technology(ies), and includes the sensor(s), a signal analysis capability, and an associated alarm indication. Often a detector is engineered in a purpose-designed physical format that may include such enabling components such as power, digital control, environmental packaging, and alternative operator interfaces.

Detection The discovery or finding of a target or non-target object. The operator is made aware of the presence of the object by means of a true alarm indication on an alarm indicator.

Detection is an active decision by the operator, or from the detector itself if it includes addition decision making algorithms (thresholding, automatic target recognition, etc.). (Adapted from CWA 14747[1])

Detection halo The circle around the actual location of a test object, within which an alarm indication is considered a true indication of detection when performing blind detection tests. The halo is typically constrained to a two-dimensional search plane, such as a wall or ground surface, immediately adjacent to, and measured from the nearest edge of, a test object. (Adapted from CWA 14747[1])

False alarm An alarm indication that is not produced by a true test object or an unintentional alarming object. A false alarm becomes a false detection when a decision is made. (Adapted from CWA 14747[1])

False detection A detection decision that is not produced by a target or non-target test object. A detection associated with an anomaly.

False test object An object or environmental perturbation not intended to be detected, that is introduced intentionally into the test site and that may generate an alarm indication. It is an item that can be representative of a non-target alarming object, i.e. clutter. (Adapted from CWA 14747[1])

Non-target object Alarming objects that generate an alarm indication for the detector, by virtue of its physical operational principles, but are not the designed-intended target objects for the detector. Metal fragments are non-target objects to metal detectors designed to detect explosive hazards.

Sensor A device that measures a physical quantity and converts it into a signal which can be read by an observer or by an instrument. A sensor is a component of a detector.

Target object Objects for which the detector was specifically designed to detect. Objects that are of interest to the operator. A landmine is the target object for a military metal detector, whereas a metal fragment is not.

Test object Object deliberately placed for testing. There are two kinds of test objects: true test object, and false test objects. (Adapted from CWA 14747[1])

True test object Alarming object that is introduced intentionally into test site in order to test the detection performance of a detector. It is an item that can be chosen to be representative of the target class which the detector is designed to detect, or it can be a simple object to be used in sensitivity measurements. (Adapted from CWA 14747[1])

True alarm indication Alarm indication caused by the presence of a target or non-target alarming object. A true alarm indication becomes a detection when a decision is made. (Adapted from CWA 14747[1])

Unintentional alarming object Objects in the test site that generate an alarm indication but were not introduced for a test.

CHAPTER 1 - INTRODUCTION

1.1 PURPOSE

To describe military search equipment testing[2] considerations in order to increase the interoperability and mutual assurance of the allied military search capability.

1.2 BACKGROUND

The current military search capability has been developed through extensive experience gained on operations in a number of theatres. It has demonstrated operational effectiveness within the Balkans, Iraq, and Afghanistan. The success of the capability has promoted the adoption of military search principles, procedures, and equipment by many NATO countries, in line with the threats now faced. Intermediate and advanced searchers must be provided with adequate tools and equipment for the task and threat, in order to mitigate risk to searchers and increase the level of assurance. As capability gaps are identified through changes in the level of threat and emerging trends, the capability must be adapted accordingly.

1.3 SCOPE

1. This publication describes technical, human factor, and operational considerations applicable to the testing of current and future search equipment, as well as providing an introduction to the application of statistics and statistical analysis to the assessment and reporting of detection technologies used in military search While this publication may be used to support testing within the context of a nation’s requirements and acquisition process, it is not designed to influence that process.
2. This publication primarily focuses on testing of specific military search equipment such as detectors and sensors. Testing of enabling equipment falls outside of the scope of this publication.
3. Chapter 2 provides general guidance and considerations regarding the trials[3]. Chapters 3, 4, and 5 describe considerations of the technical, human factor and operational aspects, respectively.

1.4 TARGET AUDIENCE

This document is aimed at personnel involved in capability development through the testing, assessment, and reporting of military search technologies. However, its contents can be relevant for other personnel involved in military search.

[2] Oxford: A procedure intended to establish the quality, performance, or reliability of something.
[3] Trials consist of a series of tests.

INTENTIONALLY BLANK

CHAPTER 2 - TRIALS

2.1 PURPOSE OF TRIALS

1. Trials, which consist of a series of tests, serve the primary purpose of verifying and validating equipment requirements. They also serve the purpose of generating user confidence in the equipment once they are issued.

2. Trial teams should ensure that trials are comprehensive, address the performance and suitability issues and are supported by appropriate technical expertise, logistical support, infrastructure and personnel.

2.2 TRIAL PLANNING

Trial teams should ensure that a trial uses structured and recognised methodologies. These should be based on set performance criteria, standards and uses representative operators, equipment, conditions and environments.

2.3 TRIAL PRINCIPLES

The design of the trial will be determined by the trial aim, objective and scope. The trial methodology and individual tests are developed and organised with resources specifically to collect data and information to answer objective questions in an effective and economical manner.

2.3.1 Clear aim

The trial aim must be clear, using NATO terminology and definitions. It must be focused on evaluating and assessing aspects of performance and/or suitability.

2.3.2 Relevance

Military search equipment is tested in order to ensure that prescribed targets can be found during specific search tasks in predetermined environments. Proper considerations should be taken to ensure that the trial results will efficiently meet the trial aim.

2.3.3 Testing independence and objectivity

Credible testing requires objectivity and independence. Trial teams should provide independent validation, free from trade-off constraints, ensuring that the equipment meets the capability requirement it was designed for.

2.3.4 Factors affecting trial design

1. The iterative trial design process requires the continual review of the following: variables, limitations and constraints.

2. The trial design process should identify variables which may affect the outcome of any test. They must be understood and completely identified so control decisions can be made to minimise or randomise their effect.

3. Every trial has implicit and explicit limitations and constraints associated with time, resources, funding, weather, availability of training areas, terrain, facilities, personnel, and safety. The effect of limitations and constraints can be moderated by flexible and creative trial design.

2.3.5 Priorities and data collection

The trial design process should establish priorities for the trial itself and confirm its objectives. The data collection should be in line with the priorities and objectives of the trial.

2.3.6 Requirement for realism

The trial design process should result in representative operators conducting operational and user tests in realistic operational and training environments. However, this should only be completed following the confirmation that the technical performance is in line with the requirements.

2.3.7 Requirement for user de-briefs

The trial design process should, where possible, include interviews, de-briefs, and questionnaires. These should correlate to one or more user requirement and be matched to performance and suitability requirements. The trial team should review previous technical and other trial de-briefs to ensure lessons learnt are understood and where appropriate are applied.

CHAPTER 3 - TECHNICAL PERFORMANCE ASPECTS

3.1 INTRODUCTION

3.1.1 Aim

The aim of technical testing is to measure the performance of a detector or sensor against a relevant target set under the influence of individual test variables with all other conditions controlled.

3.1.2 Objective

Technical performance testing provides the data and valid observations to support evidence-based decision making throughout the capability development cycle. Testing is done to a high standard that maintains the credibility and value of the results so that the work can be shared to the mutual benefit of all interested parties.

3.1.3 Scope

Technical performance testing measures and observes the functional performance characteristics of military search equipment, but not how they contribute to mission effectiveness.

3.2 GENERAL

1. A technical performance test is considered to be a scientific, objective, statistically valid observation of a measureable performance characteristic, while identifying, controlling and/or excluding all variables and externalities.

   a. **Scientific**. Scientific means that the testing procedures follow scientific principles in that they are objective, systematic, methodical, repeatable, and well-documented. Adherence to established scientific principles is critical to establishing the credibility necessary to support decision making and to provide the desired multilateral impact and value.

   b. **Objective**. Objective means that a purposeful effort is made to reduce and remove personal biases, prior beliefs, and subjective interpretation from all measurements. Objectivity influences experimental design, with an aim to create reproducible results. It also influences data analysis, interpretation, and reporting.

   c. **Statistically valid**. Statistically valid observation means that the interpretation and presentation of results are consistent with the number of measurements and the measurement precision, and that the statistical certainty is expressed along with the results.

   d. **Measurable performance characteristic**. A physical parameter that can be shown to be linked to detector performance and that can be isolated and measured.

e. **Variables and externalities**. Physical parameters or environmental characteristics that may affect detector performance.

f. An example of a simple technical test can be found in Annex A. This test gives the basic guidelines that should be followed for scientifically testing equipment. There are further considerations that must be taken in order to make it more statistically valid. A good example of a detailed technical test report can be seen in the reference TNO 2018 R10157.

2. The factors of technical testing presented in Figure 3-1 are discussed in more detail below.

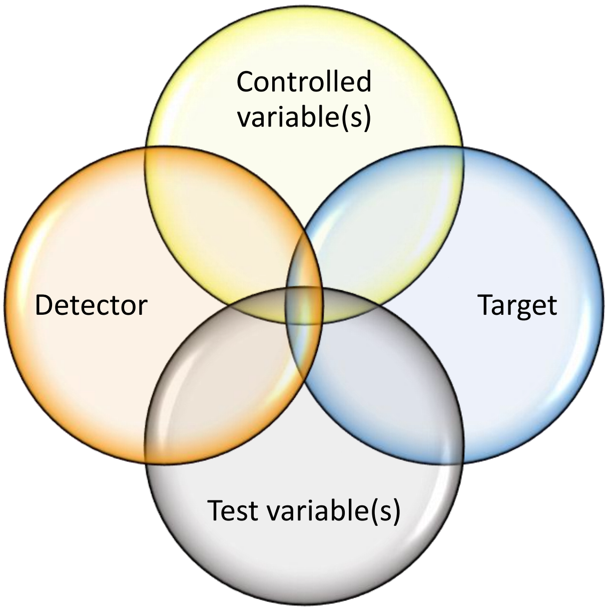


Figure 3-1: Technical testing factors.

3.2.1 Detectors

1. Fundamentally, detection of any object or materiel – be it detection of a ship at sea, or a buried explosive hazard – can be described simply as differentiating the target object from its surrounding.

2. That is, detection of an object requires the identification of a physical parameter of the target that is different from its surroundings, and then exploiting that parameter to differentiate it from its background through physical observation using a purposely designed detection system; a detector.

3. A detector is composed of:

a. one or more sensing elements;

b. components — digital or analog — to measure the magnitude of a received signal;

c. user interface, that enables user configuration of the system and presents signal information for human decision making;

d. the system may contain additional software and algorithms that analyse the measured signal to make higher-level decisions; and

e. physical housing, making it militarily useful.

4. The performance of the detector as a whole is a combination of all the these components, each of which are subject to their own operating characteristics and possibly influenced by differing environmental or operational conditions. It may or may not be possible to identify all the relevant inter-relationships, or through experimental design isolate and study all the underpinning performance characteristics, although this should be the goal where practicable.

5. Once a differentiating physical characteristic of a target is defined, a sensor is chosen to exploit that property. Sensors are often the fundamental limiting component to the operation of a detector, around which all other components are designed and optimized.

6. Sensors convert one physical observable to another more convenient physical observable, such as the conversion of magnetic fields to electrical currents that are interpreted via audio output in a metal detector. A wide variety of sensors exist, each designed to achieve a specific conversion of physical input to output observables. However, the most commonly used sensors convert inputs to electrical signals, such as current or charge, for easy digital conversion.

7. Common input signals used in military search equipment are:

a. electromagnetic (e.g. metal detector, magnetometer, radar);

b. electro-optic (e.g. closed circuit television (CCTV), ultraviolet, visible, and infrared cameras);

c. trace and vapour particulates (e.g. chemical sniffers, Raman detectors);

d. ionizing radiation (e.g. X-ray imaging, neutron imaging); or

e. acoustic (e.g. sniper locators).

8. Performance of these sensors are subject to different environmental variables and performance characteristics.

9. When packaged as a functional unit — including sensors, readouts, software, user interfaces, and physical structure — a detector may be sensitive to additional variables beyond just those of its sensing components.

10. In general, the chain of events in the operation of a detector can be described as:

a. transmission of probing signal through the air and/or soil column, or other concealing material, to the target;

b. interaction of the probing signal with the target, thereby reflecting the incident probing signal, modulating the incident probing signal, or possibly generating a new signal type altogether;

c. propagation of the target signal back to the receiving sensor; and

d. sensing, measuring, and reporting the received signal.

3.3 TECHNICAL TESTING

1. The goals, techniques, and outcomes of technical performance testing vary from the most controlled and isolated testing to the most realistic and complex environments, and could be related to the verification testing necessary at each step of the well-known Technology Readiness Levels (TRL):

   a. testing of individual system components in a highly controlled laboratory setting;

   b. testing of the integrated detector in a highly controlled laboratory setting;

   c. testing of the detector in a controlled, simplfied mission-relevant environment or setting; or

   d. testing of the detector in a complex, high fidelity mission-relevant environment.

2. The features of a credible technical performance trial are:

   a. the required performance characteristics are clearly expressed, and relevant environmental and experimental variables and parameters are identified;

   b. a scientifically valid and objective experimental procedure is described, aiming as much as possible to exclude the influence of the operator on sensor performance observation;

   c. relevant measurements are taken and recorded in a scientific and repeatable manner; and

   d. results are expressed with associated statistical confidence in a format that enables evidence-based decision making.

3.4 VARIABLES IN TECHNICAL TESTING

1. Detectors, including their sensors and other components, will be subject to a wide variety of operational and environmental variables that ultimately influence their achievable performance.

2. The following categories and technical performance factors could be considered when setting up trials to verify the technical performance of military search equipment:

   a. Detector performance factors and testing metrics:

      (1) comparison against in-service equipment;

      (2) sensor operation:

         (a) discrimination;

         (b) repeatability;

(c) limit of detection;

(d) dynamic range;

(e) probability of detection (PD);

(f) false alarm rate (FAR);

(g) sensitivity;

(h) specificity;

(i) localization;

(j) rate of advance; and

(k) variability in manufacture quality.

(3) SWaP-B (Size, Weight, Power, Bandwidth);

(4) impact/robustness;

(5) interoperability and compatibility (detector as part of the 'System');

(6) radiofrequency environment/electromagnetic interference/ electromagnetic compatibility;

(7) logistics - Reliability/Availability/Maintainability/Dependability (RAMD):

(a) open architecture;

(b) maintainability;

(c) updateability;

(d) operating endurance;

(e) storage characteristics;

(f) transport requirements; and

(g) maintenance requirements.

3. These detector performance factors are then tested against experimental variables that would likely impact required performance. When designing a technical test, select a single variable to evaluate, and then control all others as much as possible. A representative, but incomplete, list of potential variables is:

a. environment[4]:

(1) temperature:

(a) minimum;

(b) maximum;

(c) rate of change;

[4] Refer to NATO STANAG 4370

(d) temperature shock; and

(e) use & storage.

(2) humidity;

(a) maximum; and

(b) use & storage.

(3) illumination;

(a) solar load;

(b) time of day; and

(c) natural versus person-made light.

(4) vegetation:

(a) type;

(b) density;

(c) moisture content;

(d) colour; and

(e) reflective properties.

(5) soil composition:

(a) sand/clay/loam;

(b) size distribution;

(c) stratification;

(d) mineralization;

(e) water content;

(f) electromagnetic properties; and

(g) terrain and topography.

(6) infrastructure:

(a) environmental limitations (e.g. available ambient radio-frequency); and

(b) various types (urban, rural).

(7) detector employment concepts:

(a) orientation;

(b) distance to target;

(c) structural complexity; and

(d) operational mode.

3.5 TARGETS

1. The final consideration in technical testing is to define the target sets against which the detector performance will be evaluated.

2. Targets can be described as, either:

   a. **Fundamental**. Targets designed and engineered to respond specifically to the physical operating principles of the device under test (eg. the calibration target provided in a metal detector system), easily replicated or reproduced by others; or

   b. **Representative**. Targets selected to represent the real-world application of the device under test (eg. an anti-personnel landmine for metal detection), often ‘one-off’ examples that are difficult to reproduce between trial series.

3. Factors associated with the target-based evaluation may include:

   a. scenario emplacement (buried, off-route, surface laid, etc.);

   b. target-detector orientation; and

   c. placement near obstructing or interfering materials.

3.6 SAFETY AND REGULATORY CONSIDERATIONS

1. **Safety**. As most detection systems are active, in that they emit electromagnetic or ionizing radiation, additional consideration may be given to national, international, or allied regulations for safety and security:

   a. electromagnetic radiation hazard (RADHAZ);

   b. ionizing radiation regulations;

   c. laser safety;

   d. transport and storage of dangerous goods; and

   e. handling of explosive materials.

2. **Regulatory**. Legislative and regulatory considerations:

   a. spectrum management; and

   b. security of information.

3.7 ADDITIONAL PLANNING CONSIDERATIONS

1. **Facilities**. Do you have the facilities equipped to operate detection systems in a controlled and repeatable environment, utilize potentially hazardous targets, or operate complex or highly technical measurement apparatus?

2. **Qualified personnel**. Do you have the qualified personnel to operate detection systems, utilize potentially hazardous targets, or operate complex or highly technical measurement apparatus?

3. **Regulatory approval**. Have you obtained necessary regulatory approvals to operate potentially unique, novel, and/or hazardous active detection systems? Obtained approvals to acquire, transport, handle, and dispose of potentially hazardous targets?

4. **International agreements**. Is your data, analyses, and reporting affected by sensitivity and security concerns or commercial or international agreements?

3.8 RESULTS ANALYSIS AND TEST METRICS

1. Once a trial series is complete and data has been collected, further analysis is required to assess and report the performance of the device under test in a manner they relays the investigators' belief in the accuracy of the results. A discussion on experimental statics is presented in Annex B.

2. Of particular interest is the performance of the detector against a considered target set. For any detector-variable-target test scenario, the resulting observations can be categorized according to a truth table (Table 3-1).

Table 3-1: Truth table. Table entries are further explained in the text.

	Detection declared	**No Detection**
Target present	True positive	False negative
Target absent	False positive	True negative

a. **True positive** – Detector functioned as desired by detecting the target. Contributes to the detector's Probability of Detection (PD) evaluation. PD is defined as the number of detected targets divided by the total number of targets.

b. **False positive** – Detector alarmed against a non-target object. Contributes to the False Alarm Rate (FAR), which is a standard reporting metric for military search equipment as puts the observed PD in context, for example a 100% PD may result in an unacceptable FAR.

c. **True negative** – Common operating condition.

d. **False negative** – A target failed to be detected. This represents a potentially dangerous situation in operations.

3. This performance metrics are often reported through the Receiver Operating Characteristics (ROC) curve, which is described in Annex B.

CHAPTER 4 - HUMAN PERFORMANCE ASPECTS

4.1 INTRODUCTION

4.1.1 Aim

The aim of this chapter is to describe human factor (HF) trial considerations for military search equipment.

4.1.2 Objective

The primary objective of the chapter is to explain the human factor integration process and describe different human factor categories that could be considered relevant during military search equipment trials.

4.1.3 Scope

The scope of this chapter is to provide guidance on the assessment and management of human factors within Military Search equipment trials. The chapter focuses on the key areas to test, rather than the test methodology. Some human factor categories have a strong relation with the chapters on technical and operational testing.

4.2 GENERAL

1. Human factors integration (HFI) is the process by which the human component is brought together and made to work in or with a system. It is a systematic process for identifying, tracking and addressing human related considerations when testing search equipment and ensure a balance between technologies and human aspects of capability. Because the human user will have an major impact on system performance and vice versa, it is important to include human factors when testing equipment for military search purposes.
2. HFI effectively integrates the human, processes, technology and environment components, as illustrated in Figure 4-1.
3. The human component refers to the military searchpersonnel, including the organisations within which they work. The technology component refers to all of the military search equipment, hardware, software, information and materiel necessary to test the required capability. These two components are linked by organisational and management processes that include ways of working, operational tactics, techniques and procedures, and initial training before testing.

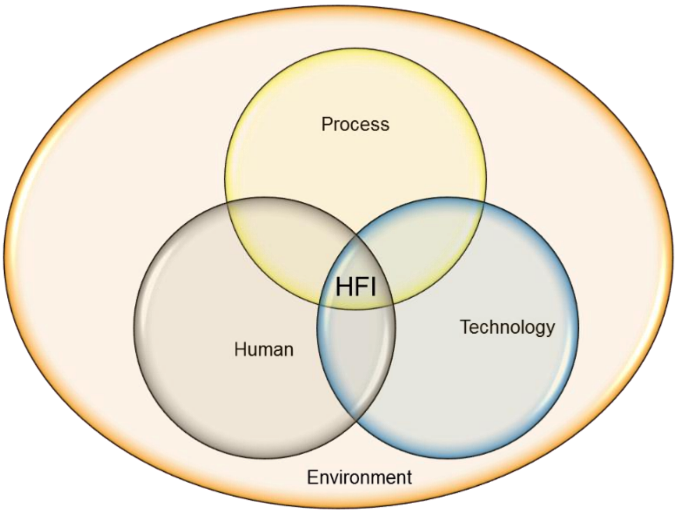


Figure 4-1: Components of HFI

4. HF categories are often judged subjectively and it can be difficult to design test procedures that allow for a quantitative assessment of HF categories. The test should be done with an average person or persons who represent the 5 to 95% percentile of a certain target audience. In order to achieve a valid test result, all aspects of the human component must be successfully integrated, taking into account the environment in which the capability will be used. The design of the tests should make best use of human capabilities (physical, cognitive, psychological and social characteristics).

5. Although many acquisition projects are concerned with the acquisition of technology (i.e. infrastructure, platforms, equipment, hardware, software), even in so-called unmanned systems, such tangible items must be operated, maintained and supported by humans. Thus whatever their nature, degree of complexity or technological sophistication, military search systems still require HF testing.

6. Failure to consider the human component of the capability can have many adverse consequences such as: increased risk of accidents and incidents, higher training costs, reduced performance and mission effectiveness, scarcity of appropriately skilled personnel, delays to the project schedule, and substantial increases in design costs.

7. To achieve the required capability, both of the human and equipment components must work together effectively and efficiently . These components are then typically linked with operational, organisational and management processes (the process component) as illustrated in Figure 4-1. Human centred design is the approach that seeks to accommodate human needs within the design of technological products or systems.

4.3 HUMAN COMPONENT CATEGORIES

4.3.1 Physical factors

1. This paragraph gives guidance on the physical factors that can be reviewed or tested during military search HF trials.

2. **Effects of equipment**

   a. **The amount of operators needed (manpower)**. This concerns the number of persons needed to operate the equipment. It includes the set up, storage and transportation, the operational use, maintenance, and if applicable, analysis of generated data.

   b. **Weight**. The weight of a certain piece of equipment has an effect on the endurance of the operator. The effects of fatigue can lead to incorrect use of the equipment leading to a decrease in operational effectiveness.

   c. **Size and form**. The size and form of a piece of equipment has impact on the ease of use. Not only can the form of the equipment limit the operator’s ability to complete their mission, but also the environments in which they can operate. Furthermore, the form or size of the equipment can create a physical burden, potentially limiting the view of the operator and impacting situational awareness (SA).

   d. **Control interface**. Can lead to difficulties in operating controls at low/high temperature.

   e. **Design**. A poor or unappealing design of equipment can lead to operators being resistant to using the equipment. Their reluctance can be based on tactical reasons or are based on aesthetics. The operators’ view on this point can be mapped by, for example, a questionnaire.

3. **Effects of other tactical, special and protective equipment**. During the execution of search tasks, the operators generally wear protective gear (e.g. personal protective equipment (PPE)) and carry other special equipment. The constraints caused by the military search equipment should not reduce the effects of protective or special equipment for the operators. Therefore military search equipment should be compatible with existing protective and tactical equipment.

4. **Effects on operator stamina**. The execution of military search tasks can be of long duration. The operator may get physically fatigued as a result of the physical burden caused by the equipment. Therefore the military search equipment should be tested by representative search operators. Where possible, recovery time from the physical burden should also be tested.

4.3.2 Psychological Factors

1. This section gives guidance on the psychological factors that can be reviewed or tested during military search equipment trials.

2. **Effects of equipment**

   a. **Mental workload** (cognitive burden). The amount of information which is presented to the operator has to be taken into consideration. Cognitive overload may affect proper use of the equipment, execution of the search procedure and the SA of the operator. Therefore, it is important to review the degree in which the operator needs to interpret information, maintain SA of the environment, perform detection and operate the equipment at the same time.

   b. **Maximum time of operation**. Depending on the cognitive workload required by the equipment, an operator may reach their cognitive limit. The time the operator is able to operate the system without losing focus should be considered.

   c. **Recovery time** for operator. This aspect has closely aligned with maximum time of operation. In the trial, the amount of time needed for the operator to be well enough rested on cognitive level should be assessed.

   d. **Control interface**. This concerns the way the system is controlled and in which way the settings have to be adjusted during the start-up and operation of the equipment. The following aspects should be reviewed:

      (1) intuitivity: the ease of interpretation of the controls;

      (2) menu structure: clarity and ease of interpretation of the system’s menu; and

      (3) time to change settings: the time an average operator needs for adjusting or changing settings.

   e. **Information interface**. This concerns the way the information is presented to the operator and if the operator is able to interpret the information under all conditions.

      (1) Reduction of interpretability under harsh environmental conditions (sunlight, rain, dust, low/high temperature). There can be sun reflections, wet, windy or dirty conditions which make it hard or impossible to interpret the information presented by the equipment.

      (2) Type of presentation interface. Several types of interfaces can be used for the presentation of the information to the operator. Depending on the type of sensor, environment, and task, the type of presentation can have benefits or shortfalls.

         (a) **Visual**. Information can be presented visually e.g. LEDs or monitors. Aspects that can be assessed are:

            i. Operability in day and night conditions. Due to over-exposure of light, the operator can temporarily lose his eyesight during night time.

ii. Lights. The operator has to be able to recognize and interpret the visual signal given by the equipment, e.g. different colours, blinking, etc. This is especially true when the detector is able to indicate more than one parameter, e.g. in dual sensors.

iii. Display. A display can present the information in different ways, for instance, with a plot or image in full colour or grey scale. The operator should be able to interpret the displayed information under all conditions. A downfall of a large display can be a loss of SA because the operator will be more inclined to watch their display instead of the environment.

(b) **Audio**. Some outputs are presented with audio signals. The operator has to be able to recognize and interpret the signals under all conditions. This is especially true when the detector is able to detect or indicate more than one parameter.

(c) **Tactile**. Some outputs are presented with vibration signals. The operator has to be able to recognize and interpret the signals under all conditions. This is especially true when the detector is able to detect or indicate more than one parameter.

3. **Training**. For new equipment, especially when new technology is involved, additional training is required. The amount of training needed depends on several factors such as experience, but also the complexity of the detector control and technology. It is important to determine the training effort needed because it can have operational impacts. It is suggested to review the different categories of personnel (operator, instructor and maintenance) involved in the usage of the system, because they all have their own particulars. To consider:

a. need for a prior knowledge. Does the operator need specific knowledge (e.g. about sensor technique) to be able to work with the equipment?;

b. amount of initial training needed; and

c. long-term training needed to maintain skill and prevent skill fade.

4.3.3 Safety and health hazards

1. This section gives guidance on the safety and health hazards that can be assessed during military search equipment trials in relation to HF. Risk of injury or exposure to health hazards in the long and short term caused by operating the system should be reviewed. Possible sources of injury can come from the equipment itself (e.g. if the system malfunctions, radiation, toxic fumes) or being caused by using the equipment (e.g. repetitive strain injury, operator error).

a. **Radiation**. Equipment involving radiation poses a health hazard for the operator. In some cases, the maximum exposure time within national radiation legislation should be measured. Radiation legislation can cause a restriction in the use of the equipment and thus form a limitation to the execution of military search tasks.

b. **Workload, movement and ergonomics**. The weight and size of a piece of equipment in combination with the movement required to operate it can cause injuries to the operator. It is recommend that specialists, such as physiotherapists, are involved in the trials so they are able to measure, estimate or predict the effects of the workload on the operator.

CHAPTER 5 - OPERATIONAL TESTING

5.1 INTRODUCTION

5.1.1 Aim

The aim of this chapter is to offer guidance on operational trial considerations for military search equipment, and to assist in the assessment of military search equipment during scenario-based search tasks.

5.1.2 Objective

User trials are designed to assess operational useability and feasibility. Additionaly, they provide the means to assist in the development of equipment. Trials give advice and direction on the support required to move the equipment through production status to acceptability for operational use as well as supportability in-service.

5.1.3 Scope

Operational testing is to ensure equipment is suitable for use in the operational environments where it is expected to be used. Operational tests must be conducted in a manner that is representative of the operational environment and conditions, using trained troops.

5.2 OPERATIONAL PERFORMANCE CRITERIA

The following factors address how well the system performs during critical missions, or during the tasks it was designed to do.

5.2.1 Mission performance

This addresses a system's ability to meet the stated operational capability requirements under anticipated operational conditions. The Battlefield Mission (BFM) is a mission schedule that specifies a typical representative operational requirement over a set timeframe. These normally address key functions, e.g., detect, recognize, identify, engage, process information. It requires information to be collected on a system's capability to perform its intended role and tasks when operated by the end user under field conditions.

5.2.2 Survivability and vulnerability

This addresses a system's ability or likelihood to avoid being rendered ineffective whilst conducting a BFM. Test measures are normally expressed in terms of task requirements and exposure times to determine the ease of use during operations. Other measures determine the extent of damage, given an enemy engagement or task. It requires information to be collected on how well the system avoids enemy detection and the level of damage during operations.

5.2.3 Suitability Issues

These issues address the significance and impact of a system's demands for support in order to remain operationally effective. It examines what adjustments must be made to the logistics and training system to service, maintain or operate the new item without disruption of service or support to existing systems.

5.2.4 Reliability, availability and maintainability (RAM)

1. This addresses whether new or modified systems can be depended upon to perform where and when required. Testing determines what appropriate user support is required to keep the item functional during training and operations when employed by the intended users.

    a. Reliability deals with the assurance that a system will not encounter an unacceptable number of failures during operation, generally expressed as Mean Time between Failure (MTBF).

    b. Availability is a measure of whether an item is ready to be used.

    c. Maintainability deals with the ease of repairing or replacing a system component that failed. It accounts for the time to diagnose, repair or replace and test. It is generally expressed in Mean Time To Repair (MTTR) the system to an operating condition.

2. RAM testing requires information to be collected on the system's capability to be ready to perform its missions, the reasons for the periods when the system is not ready or degraded, and the requirements to return the system to a ready status. This requires data collection, analysis and evaluation over time, and support may be required.

5.2.5 Logistics supportability

This addresses what is required to assure that a new or modified system can be supported fully when fielded. Logistics supportability deals with the impact of providing maintenance and operating support. Maintenance concerns include repair teams, procedures, and spare parts supply. Operating support considers expendable items (POL, batteries, swabs, filters, etc.). Transportability and deployability deal with the ability to move the system to a theatre of operations and move within the theatre. Logistics supportability testing requires information to be collected on the personnel, materiel and documentation required to support the system. Operational testing dealing with this issue is related to field logistics, maintenance and repair by the user organisation. Information must be collected over time and the evaluation of such data may require support.

5.2.6 Training

This addresses the capability for, and the adequacy of training for, user personnel to operate the system. It should also address the training requirement for support personnel. It includes some consideration of HF issues and requires the collection of information on the adequacy of training programs and the capability of personnel to be trained on the system. Capability directorates, schools, SMEs and R&D staff may be consulted.

5.2.7 Employability

This deals with organisational, doctrinal and tactical issues to examine if the current procedures, organizations and tactics are compatible and effective with the proposed equipment. Organisation addresses the distribution of the equipment, its maintenance and support. Doctrine addresses the suitability of planned or current doctrine to effectively employ the system. Tactics addresses drills and procedures for employing the system.

5.2.8 Compatibility and interoperability

Compatibility addresses how well the system operates or interacts with other battlefield systems, including the extent to which it does not interfere with them. It examines the integration of the item with the systems with which it is expected to operate. It could also examine interoperability with allied systems.

5.2.9 Software

This addresses the adequacy of software within the system. It requires the collection of information about the performance and problems encountered during software operation. User testing concerns utility, ease of use, responsiveness, safety, effects on system performance, ease of changing or updating the software, and skill fade when not using the system continually.

5.2.10 Security

This issue addresses maintenance of the integrity of the system’s security classification and preventing information about its characteristics, performance, technologies and reliability from inappropriate disclosure. It examines the ease of foreign technical intelligence to gain access to, or information about, the system. It also concerns the control of training, information and manuals, and physical security.

5.2.11 Acceptability

This issue addresses the typical user’s satisfaction with the system and its ease of use and operation while performing battlefield tasks. It also addresses their confidence in the system’s performance and reliability.

5.3 STANDARDS

Operational trials conduct a comprehensive evaluation of the total system. However, senior decision makers should address the critical or key requirements to approve the system for various levels of acceptance related to whether or not the equipment provides the required capability or has the potential to do so. This requires prioritising issues to ensure that trial design focuses on the key issues so that, if the trial is curtailed for any reason, useful information is still obtained. Issues are categorised as key user requirements, key system requirements, and other requirements.

5.3.1 Key user requirements (KUR) and key system requirements (KSR)

1. KURs deal with the total system's requirement to be ready for, and sustained, during operations. KURs must be achieved to meet the stated operational capability. Each KUR should have a threshold, objective and measure of effect. The threshold is the minimal requirement for the capability; the objective is the desired requirement. The requirement must be operationally relevant (focused on the mission/role) and address performance effectiveness, suitability and safety of the equipment. Factors such as the relevant doctrine, tactics, climatic environments, operating theatres and restrictions on the exportation of specific equipment should be examined.

2. KSRs are similar to KURs but are more focused on specific performance, suitability characteristics and related specifications. This includes management issues such as costs, risks and timelines.

3. KURs and KSRs focus on the following:

   a. **Operational effectiveness** is focused on how well the equipment performs the missions and tasks it has been designed to do.

      (1) Performing critical missions or tasks. The equipment must be assessed on how it performs the critical tasks that it has been designed to do. Any failure here should be viewed as a critical failure as the equipment cannot achieve a key requirement.

      (2) Durability for mission or tasks. The equipment should be robust enough to survive the battlefield mission and be able to be re-tasked to another mission. Equipment should be durable enough to survive the mission in all the environments in which it will be employed.

      (3) Interoperability with other battlefield systems. The equipment must be compatible with other equipment specific to the mission; such as personal equipment, ECM, remote control vehicles, other detectors, and vehicles. Strong consideration should be given to interoperability with other nations' equipment.

      (4) Strategic and tactical mobility. The ability of the equipment to be transported into and around a theatre of operations needs to be assessed. Transportation within the theatre of operations should consider air, land and water-based vehicles as well as being carried by the operator. Consideration for specialist users such as airborne and commando forces should be included to ensure the equipment is suitable.

   b. **Operational suitability** assesses the support the equipment requires to remain operationally effective. These support requirements should also be considered against the wider impact on the mission. If a piece of equipment requires significant support, this is likely to have an impact wider than just the unit using the equipment.

(1) Ability of operators to use and maintain their proficiency on the system. Consideration should be given to how complex the system is and how much training is required by the user to maintain their proficiency. The more complex the system, the greater the cognitive burden for the operator. Where possible, technology should be used to reduce the cognitive burden on the operator to facilitate greater endurance.

(2) Ability to maintain the system. The ability of the user and equipment support teams to maintain the systems needs to be assessed. A judgement will be required based on how much maintenance is required and the effect this will have on system availability.

5.3.2 Other requirements (OR)

1. ORs deal with the performance of various components or elements of operational suitability and effectiveness. ORs supplement the critical issues to ensure a thorough investigation of an item. In some cases, they address the differences between the user requirements and system requirements, non-performance criteria or requirements, user acceptability, training and other areas which should be examined to ensure that a cost-effective, supportable and suitable system is acquired. Often, it is these important additional issues which determine the acceptance of an item, particularly if a number of competing items are being evaluated. Typical ORs include operational effectiveness and operational suitability.

2. **Operational effectiveness**.

a. **Software compatibility**. The equipment software should be tested for compatibility with the other systems that the equipment will interact with on the battlefield. The key areas for interaction and compatibility should have been articulated as a system requirement and first be tested during technical trials. The operational trial should confirm this compatibility and suitability.

b. **Personnel capabilities**. Consideration should be given to the type and amount of training required to develop competence with the equipment. Training for the usage and maintenance of the equipment should be a factor for consideration during the procurement process, the operational trial should be used to confirm requirements.

c. **Organisational equipment support structure**. The logistical support and sustainment of the equipment should be assessed to ensure organisational readiness for the equipment. The organisation should be suitably structured for the equipment support. Deficiencies in this area should be seriously considered due to the impact on sustainability of the equipment. People, training and equipment required to operate and maintain the equipment need to have been considered during the procurement process.

d. **Vulnerability and survivability of sub-systems**. Sub-systems should be assessed to ensure they are suitably robust and survivable. Sub-systems must not be a failure point within the overall system. Integration of sub-systems should also be factored in and assessed during operational trials.

e. **The adequacy of the doctrine and tactical procedures and techniques**. Doctrine and policy should have been written to support the capability concurrent to the procurement of the equipment. This should be tested during the trials with observations fed back into doctrine and policy prior to equipment being accepted into service.

3. **Operational suitability**.

a. **Safety**. The equipment should be safe to use for the user and those operating near the equipment. The impact of personal protective equipment should be assessed with regard to the degradation of the operator's performance and ability to complete the mission.

b. **Human factors integration/human machine interface (HFI/HMI)**. HFI trials have been discussed in Chapter 4. Any issues found during these trials could be reassessed during operational trials.

c. **Reliability, availability, and maintainability (RAM)**. The availability of the capability is key and observations on the equipment's reliability and requirement for maintenance should be considered.

d. **Manpower issues such as endurance and skills**. Assessment of the physical burden for the operator should be undertaken. This assessment will develop understanding of the effect the equipment has on the operators physical and mental endurance.

e. **Logistics supportability**. Considerations should be given to the requirement for logistic support to maintain the equipment. Logistic support to the equipment should have been considered during the equipment procurement process and any key requirements should be assessed.

5.4 OPERATIONAL ENVIRONMENTS

Operational trials will need to be conducted in the environments that the equipment will be used operationally. These trials should be conducted in a variety of geographic locations while also covering the following key environments: subterranean, urban, and rural. The environments should be set up to be as realistic as possible so that the equipment can be tested as robustly as possible. User and system requirements which were set early in the procurement should be included in the trial plans.

ANNEX A - TECHNICAL TEST REPORT EXAMPLE

A.1 INTRODUCTION

This annex gives users an overview of the considerations that should be thought about in order to conduct a technical test. Below is an example of a simple technical test report.

A.1.1 TEST PARAMETERS

Before designing any test, the aim should be clearly defined. The aim needs to be translated into measurable parameters. For the example below, the aim was to establish the discrimination performance of a certain metal detector for buried 20 mm projectiles from different buried metal objects. The test parameters were if the object found was a 20 mm projectile or not, and how accurate the perceived point of detection was from the centre of the object.

A.1.2 VARIABLES

1. During a test, variables should be controlled and recorded as much as possible. For a detailed list of variables, see paragraph 3.4.3.
2. For the example below, the variables were the air temperature and the soil moisture. Due to the test being conducted outdoors, the variables could not be controlled but were recorded. A way to control these variables is to do the test indoors.

A.1.3 TARGETS

Define the target sets against which the detector performance will be evaluated. For the example below, the target set were 20 mm projectiles at a set depth.

A.1.4 DETECTORS

1. Detectors should be used in line with the manufacturers directions. To make the test valid, detectors must remain on a constant setting and used in the same manner each time they are used to be able to compare the results. The recording of these settings is important for further analysis.
2. For the example below, one detector type, however several samples, were used by multiple operators. The detector settings were kept the same for all operators, and all operators were instructed to carry out the task in the same way.

A.1.5 PERSONNEL

A.1.5.1 Operators

Multiple operators should be used to limit the personal bias. All operators should be trained to the same standard prior to the test taking place. For the example below, 10 operators were used to carry out the task.

A.1.5.2 Neutral personnel

The test should be coordinated and monitored by a neutral test leader. Neutral personnel should ensure the task is carried out as instructed and all recordings are objective. For the example below, one neutral person was responsible for this task for the duration of the test.

A.1.5.3 Documentation

Below is an example of documentation to record a basic test.

DISCRIMINATION TEST EXAMPLE

The test administrator reads the info for the operator conducting the test.

Test setup:

Buried 20mm projectile in sand contaminated with other metal objects.

The test setup is explained to the operator.

Procedure:

Search lane from start to end (show start / end position).

The task is to discriminate between the 20 mm projectile and other metal objects in the ground.

Mark perceived centre of located targets (20 mm projectile) with red marker (show red marker).
Mark perceived centre of other objects with blue marker (show blue marker)

The actual procedure is given to the operator. The procedure should clearly state what the operator needs to do to execute the test.

Facility:

Target: 20mm projectile is buried in calibration square A (show A).

Soil compensation site is provided in calibration square B if needed (show B).

For this example, the operator had the possibility of doing a soil compensation on the detector if required (test square B). He also could test the detector output against the target in a known location (test square A).
This is not mandatory for every test, but the added value of this possibility should be examined during test design.

Time to complete:

Maximum amount of time for assignment is 30min, report to test administrator when assignment is completed.

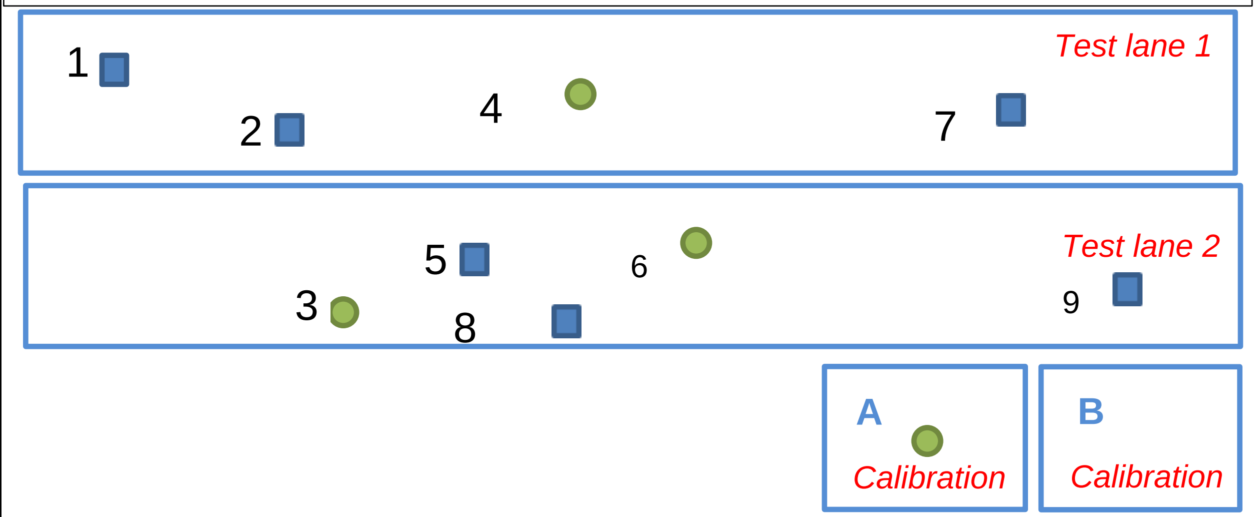
Time should only be considered if there is an optimum rate for the use of the equipment.

DISCRIMINATION TEST EXAMPLE	Location:	
	Air temp(^{o}C):	*All variables should be recorded in the report.*
	Soil moisture (%):	
	Company:	*Information regarding the operator and detector should be recorded in the report.*
	Sensor serial nr.:	
	Operator:	
	Time (min):	

Detector startup/setup: *Any remarks the test administrator notices regarding the startup of the detector should be mentioned here. These remarks can be used later in the analysis of the data.*
Soil compensation and detector settings: *Any information the test administrator notices during the soil compensation or about the detector settings should be mentioned here. These remarks can be used later in the analysis of the data.*
Search: *Any remarks about the execution of the assignment should be mentioned here. These remarks can be used later in the analysis of the data.*
Localisation / Marking: *Any remarks on the localisation or marking of the found objects should be mentioned here.* *These remarks can be used later in the analysis of the data.* *It is important not do disturb the operator nor the markings placed.*

Results: Test administrator only!

Discrimination test example



Object		Located			Deviation	
Number	Type	Yes	Marker		Positioning Error (cm)	Comments
			O	T		
1	*Coin*	*X*	*X*		*0*	
2						
3	*Describe t he buried objects here*	*Mark if the object was located*	*Mark if the object was discriminated as the target (red marker) or not (blue marker)*		*Write the distance between the perceived centre of the object and the actual centre of the object here*	*Any comments regarding the detection of the object can be written here*
4						
5						
6						
7						
8						
9						

Number of false alarms:	*Record the number and locations of false alarms given by the detector (see paragraph 3.8 on false alarms)*

Test Administrator

DTG (DDHHMM_MONYY)

INTENTIONALLY BLANK

ANNEX B - ANALYSIS & REPORTING

CONTENTS

B.1 INTRODUCTION

1. This annex is intended to provide an introduction to the application of statistics and statistical analysis to the assessment and reporting of detection technologies used in military search.

2. Statistics is a rich and deep field, necessitating significant study to master. However, a basic understanding of the concepts and language used in statistical analysis can be understood and applied by anyone responsible for the technical evaluation of detection technologies used in military search. This knowledge will enable technical authorities to share and understand the significance of their observations and, perhaps more importantly, express the limits of the inferences that should be drawn from them. What weight should a single observation of a successful detection be given by a technical authority responsible for capability development? Or, conversely, how best can a study's author relay their belief in the credibility and accuracy of their results so as to best inform the broader community of stakeholders and partners?

3. This annex intends to familiarize the reader with the basic concepts and language of statistical analysis using examples common to the technical evaluation of detection technologies used in military search, and as such drives directly to specific discussions and concepts while bypassing other important topics that would be considered in a more thorough introduction.

4. This introduction does not purport to be thorough, but rather to serve as a reference for the interested reader; to introduce the common language and concepts used within the community, and provide suggestions on how to both report their results to the community as well as understand the implications of the results they receive from others. When appropriate, related statistical concepts and terminology not covered in this introduction will be highlighted in italics for the reader to pursue, as required.

5. For those unfamiliar or uncomfortable with statistical analysis, the most significant recommendation from this introduction is to thoroughly explain your analysis and assumptions; explain the experimental procedures, any assumptions made, the data analysis, error analysis, and resulting inferences as thoroughly as possible with the intent of providing the reader with the ability to interpret for themselves the value of the repotred observations and recommendations..

B.2 STATISTICS AND PROBABILITY APPLIED TO MILITARY SEARCH

6. Detection within military search is principally concerned with locating, confirming, and identifying target objects within search missions. Performance of detection technologies in these tasks can be described in the language of Key Performance Indicators (KPIs), which are values that can be measured and quantified.

7. KPIs for detection technologies may include those directly observed, such as:

   a. depth of detection;

   b. range of detection; or,

   c. sensitivity to electromagnetic noise.

8. They may also be inferential KPIs that speak more to implied performance in operational scenarios, along with stated or unstated assumptions of the operational environment, such as:

   a. probability of detection (PD); or,

   b. the false alarm rate (FAR).

9. All these values are statistical in nature, in that they are expressions of performance based on meaningful patterns that emerge out of multiple individual observations. They become important when these objective and repeatable measures of performance are used as proxies, or indicators, for operational performance in order to inform capability development.

10. While the full machinery of statistical and probabilistic theory can be brought to bear on these KPIs, thankfully, the majority of those relevant to the assessment and reporting of military search technologies can be understood in a very straightforward manner and do not require that the technical authority to delve deeply into statistical theory in order to present and understand their observations sufficiently to influence capability development decisions. That said, in simplicity lies the risk of misinterpreting or misrepresenting one’s observations, which is minimized by the additional effort of error analysis.

B.2.1 BACKGROUND

“*Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena*.” [4]

1. Statistical analysis begins with observation and counting; eg. observing the detection, or not, of a target object in a particular environment or scenario, and counting how many times that occurs for each attempt.

2. *Data* is the set of results, or the *sample*, obtained from a series of observations. They form the set of measurements from which the investigator intents to extract meaningful information. Data is *quantitative* if it can be obtained by a quantifiable measurement process and represented by numbers – for example, the maximum range of detection – or *qualitative* if otherwise, such as an end-user's appreciation for the ergonomics of a design. The former is amenable to mathematical analysis, and will be the focus of this introduction.

3. Data may be *discrete* (represented by integers, eg. the number of bars on a detector LED display), or *continuous* (represented by real numbers, eg. repeated maximal distances of detection). Samples sets of such data form the basis for almost all relevant analysis concerning the performance of military search detection technologies.

B.2.2 NOMENCLATURE AND NOTATION

1. Some convenient mathematical notation will aid further discussions:

   a. Factorial, !, is the descending product of a series:

      (1) 5! = 5 x 4 x 3 x 2 x 1 = 120 ; where,

      (2) 0! = 1, is defined.

   b. Ellipses, … . Depending on context, are interpreted as:

      (1) the natural extension of a series, 1, … , 4 = 1, 2, 3, 4; or,

      (2) a simple abbreviation for a long list of values, {1.3, 2.4, … , 1.7, 3.2}.

   c. $\sum_{i=1}^{n} x_i$, the *summation* of a set of *n* values, iterated over $i = 1, \ldots, n$:

      (1) $\sum_{i=1}^{n} x_i \;= x_1 + x_2 + \cdots + x_n$

   d. $\{x_i\}_{i=1}^{\mathrm{n}} = \{x_1, \ldots, x_n\}$ denotes a set of *n* values, often simplified as $\{x_i\}$.

2. More specifically to statistics and probability, if *X* is a *random variable* — that is, *X* represents the possible outcomes observed from a random process, such as the measurement of the maximum detection distances of a detection technology to a target — then,

   a. $\overline{X}$, is the *mean*;

   b. $\langle X \rangle$, is the *expected value*; and,

   c. $\hat{X}$, is the *estimate* of *X*.

3. The *expectation value* and *estimate* of *X* are related to the underlying *probability distribution* governing the outcomes of the random variable, which will be discussed in more detail below.

B.2.3 STATISTICS

1. Let $\{x_i\} = \{x_1, ..., x_n\}$ denote a set of data *n* values collected in a test series. There are a number of data *statistics* that can be used to describe this set, the most common being:

   a. $Min\{x_i\}$, the *minimum* value of a set of values;

   b. $Max\{x_i\}$, the *maximum* value of a set of values;

   c. $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the *arithmetic mean* of a set of *n* values;

   d. $Var\{x_i\} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\ \right)^2$, is the *variance* of the set of values; and,

   e. $\sigma = \sqrt{Var\{x_i\}}$, the *standard deviation*.

2. The mean, $\overline{x}$, of a data set is used to present the average value from the otherwise inevitable variability observed in repeated measurements of a physical parameter. When measuring, for example, the maximal detection range of a metal detector to a target, $\overline{x}$ presents a useful summary of those measurements, when all experimental conditions are held constant.

3. The variance of the set, σ^2, relays a measure of the spread of the data set around its mean value, $\overline{x}$. A small variance relative to the magnitude of $\overline{x}$ results when the data differs minimally from its mean value, so in some sense imbues a confidence that the mean represents the true and repeatable value of that physical parameter. A larger variance may indicate that either the physical processes involved are highly variable, or, as is often the case in scenarios relevant to this discussion, the experimental technique is influenced by unaccounted-for external variables (temperature fluctuations, operator interpretation, etc.). In either case, a large variance would imply less confidence that the resulting $\overline{x}$ represents the true value parameter being measured and thus relays to the reader some measure of the appropriate belief in the results.

4. Before moving to a discussion on probability it will be useful to first introduce the histogram, as it is a very useful tool in data analysis.

B.2.3.1 Histogram

1. The *histogram* provides a method of presenting data based on the frequency of occurrence in a data set, and is applicable to both continuous and discrete data.

2. A histogram is defined by dividing the range of possible measurement outcomes into multiple intervals, or bins, with each bin having a lower and upper bound (LB and UB), so that each observation x_i can be assigned to a specific bin satisfying $x_{LB} < x_i < x_{UB}$, and the bin value subsequently incremented by one count. The bins, in effect, record the number of parameter observations that fall in the range defined by that interval and thus the histogram provides a rough estimation of the underlying probability distribution, which will become more meaningful later in this review.

3. For discrete variables the bins are typically the natural unit of enumeration, most often integers. For example, the number of landmines detected in a test lane of 100 targets. For continuous variables the number of bins, and thus the interval range per bin, is arbitrary, and is usually selected for ease of presentation or analysis.

4. Consider an example. Suppose the maximum detection distance of a detection technology to a specific target is repeated *n* = 50 times. The resulting measurements are {31.3 cm, 39.2 cm, …, 24.9 cm, 29.9 cm}, which is found to have a mean $\bar{x}$ = 30.1 cm and a standard deviation σ = 4.3 cm. These values are useful to report, but they do not relay all the information in the data.

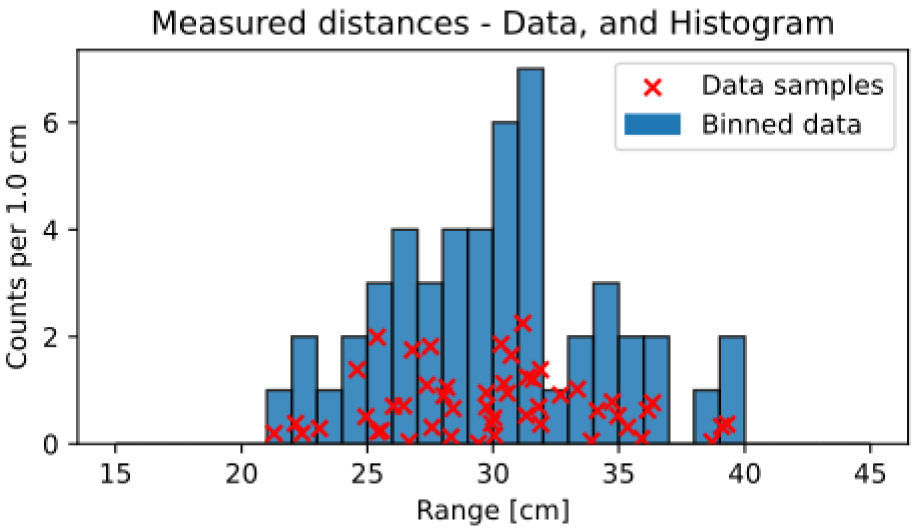


Figure B-1: In this example, raw data observations from the measured maximum detection distance to a target are depicted in red crosses. This data is naturally one dimensional, so Y-axis variation has been added for clarity. By dividing the range into 1 cm intervals, and assigning each data observation to unique interval, a histogram is constructed (blue) that begins to represent the underlying probability distribution from which the measurements were made.

1. However, consider Figure B-1, where these raw data observations are depicted in red crosses. By dividing the range into 1 cm intervals, and assigning each data observation to unique interval, a histogram is constructed (in blue). The counts per bin begin to provide a visualization of the underlying probability distribution that governed the resulting observations. As it is often that underlying distribution that is of interest, a histogram representation is often quite useful.

B.2.4 PROBABILITY

1. In addition to statistics, another concept of significant important to the assessment and reporting of detection technologies used in military search is *probability*; the likelihood of an event being observed.

2. Probability is likely familiar to most readers, as it encountered on a daily basis. The probability of rain in the forecast, or the probability or a candidate being elected, are both intuitively understood to relay some likelihood of an event occurring. Given the previous example, one could ask "What is the probability that when used on operations the detection range to the target of interest will be greater than 50 cm, or less than 20 cm?"

3. The philosophical underpinnings and interpretations of probability are deep and can lead to significant insights regarding the limits of detection technologies, such as sensitivity or minimal levels of detection, but for the most direct applications such as PD and FAR the simplest concepts suffice.

B.2.4.1 Mathematical Probability

1. Probability can be described as the numerical description of how likely an event is to occur, and is represented by number between 0 and 1 where 0 indicates impossibility of an event occurring, and 1 indicates certainty.

2. Let *X* be a random variable, or measurement, and $S = \{E_1, E_2, \dots E_n\}$, be the set of *n* possible experimental outcomes, or *events.* Events E_i and E_j are referred to as *mutually exclusive* if it is impossible for them to occur at the same time (in one coin toss one can observe either heads, E_H, or tails, E_T, but not both).

3. For *n* mutually exclusive events, $S = \{E_1, E_2, \dots, E_n\}$, for each possible outcome $X = E_i$ there is a probability $P(X = E_i)$ that satisfies,

   a. $0 \leq P(X = E_i) \leq 1$;

   b. $P\big(X = E_i \ or \ E_j\big) = P(X = E_i) + P\big(X = E_j\big)$, for $i \neq j$; and,

   c. $P(X \ in \ S) = \sum_{i=1}^{n} P(X = E_i) = 1$.

4. This is simplified considerably when one considers detection tests, where either a target is detected, *X* = *d*, or not detected, '*X* = not *d*'. Clearly, being detected or not detected are mutually exclusive events, thus,

   a. $0 \leq P(X = d), \ P(X = not \ d) \leq 1$;

   b. $P(X = d \ or \ not \ d) = P(X = d) + P(X = not \ d)$; and,

   c. $P(X = d) + P(X = not \ d) = 1$.

5. If $P(X = d) = p$, then this last equation implies $P(X = not \ d) = 1 - p$. This will be recognizable in the binomial distribution discussion below.

6. More complicated scenarios may occur, for example, when considering the assessment of target classification algorithms that attempt to classify a target from a list of possible threats. In this case the decision algorithms can be complex and the results may not be mutually exclusive; "The observation is consistent with a buried landmine OR a rock OR a piece of wood." A more detailed probabilistic assessment would be required that accounts for *correlations* in the data, which is beyond the scope of this introduction.

B.2.4.2 Frequentist Probability

1. This is the most common understanding of probability. Consider an experiment performed *n* times, and a certain outcome, *d*, occurs in *m* of these tests. As the number of tests performed becomes very large, $n \to \infty$, then the probability of outcome *d* tends toward a limit defined as $P(d) = m/n$.

2. This definition of probability suffices for the majority of tasks related to the assessment and reporting of detection technologies used in military search; "In a test lane of 100 identically buried landmine targets, 35 were detected. P(detection) = 35/100 = 35%."

3. However its application in real-world analyses requires a very important caveat; the number of tests performed, *n*, will not often tend towards infinity and may in fact be very small, or even singular in the case of expensive tests. Establishing confidence in the inferences that can be drawn from a small data set will be discussed throughout this introduction.

B.2.4.3 Bayesian Probability

1. As mentioned above the frequentist interpretation of probability largely suffices for the task at hand, however, one crucial nuance is worth discussion, made all the more important because this introduction is written for the statistical novice. The frequentist approach implicitly assumes that there exists a true underlying probability of an event, defined as the limit of a ratio as the number of observations tend to infinity, independent of the experiment making those measurements. That is, it assumes the probability is an *objective* quantity that exists independent of the investigator. An alternative view is presented by Bayesian probability theory, where the observer is considered an irremovable part of the observation, and that the observed probability is in essence *subjective* in nature.

2. The Bayesian interpretation asks one to first consider what one hypothesizes before a test is conducted, known as *a priori* hypothesis *H*, and how that hypothesis would be modified *a posteriori* through observed data *D*. Mathematically, this is expressed as,

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

3. This is read as an expression of the probability of hypothesis *H* being true given the observed data *D*, denoted $P(H|D)$, which is calculated as the probability of observing data *D* given hypothesis *H*, or $P(D|H)$*,* times the *a priori* probability of hypothesis H, *P(H)*, normalized by *P(D)*, which is the prior probability of seeing the data D under all mutually exclusive hypotheses.

4. This can be complex to apply in general, but its utility becomes more apparent when considering the assessment of Automatic Target Recognition (ATR) algorithms incorporated into some search technologies, which aim to discern a *Target* class from observed *Data,*

$$P(Target|Data) = \frac{P(Data|Target)P(Target)}{P(Data)} .$$

5. Here, the Bayesian probability is read as the probability of there being a specific target present given the observed data, which is calculated as a function of the probability of observing that data for a given target, $P(Data|Target)$, times the prior probability of that target being present in that area, $P(Target)$, normalized by *P(Data)* as the prior probability of seeing that specific data under all mutually exclusive hypotheses. Here, the $P(Target)$ and $P(Data|Target)$ terms that expressly capture the knowledge, experiences, and expectations of the military searcher before a test is conducted, and $P(Target|Data)$ shows how that belief is modified and updated by the information, *D*, that the search technology provides.

6. The Bayesian interpretation is supported by deep philosophical arguments, but in essence it demands that the investigator explicitly recognize that complete objectivity in a test is unattainable and only subjective observations are possible. While the ultimate goal of an experimental design is to completely isolate the test variables and remove the observer from influence, this is in practice very difficult and the Bayesian interpretation reminds us of this.

7. Due to its ease of application the remainder of this introduction will utilize the frequentist interpretation of probability, but the Bayesian interpretation reminds us that all observations are subjective and the investigator can only try to quantify the magnitude of that influence in part through a credible and thoroughly described experimental procedure and error analysis.

B.2.4.4 Probability Distributions

1. A *probability distribution* is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. The shape of Figure B-1 hinted at the underlying probability distribution that lead to that particular set of observations.

2. Fortunately there are a variety of analytic (mathematically-described) distributions available that can be shown to be applicable to a variety of testing scenarios based on particulars of the underlying processes and potential outcomes of an experiment [1] [2] [3].

3. It is important to note, however, that in regards to the testing and assessment of military search equipment one will most often calculate an apparent probability distribution derived from direct observation of an experimental outcomes. From this observed probability distribution, analytic probability distributions are then useful in that they allow one to infer the future likelihood of observing an event. That is, applying the results of a trial series to the anticipated performance in future scenarios. Probability distributions do not affect how one reports their observations but, importantly for this discussion, they do provide an understanding of how to estimate the errors associated with the reported observations.

4. Three distributions are of particular relevance to the assessment of military search technologies: Binomial, Poisson, and Normal.

B.2.4.4.1 Binomial

1. The *binomial distribution* describes processes that have two possible outcomes; for example, flipping a coin, or, more relevant here, 'detect' or 'not detect' when testing a detection technology against a specific target.

2. Let *X* be a binomial random variable, denoted $X \sim \mathcal{B}(p,n)$, If we denote the probability of a detection event, *d*, as *p* = *P(X=d)*, and we repeat the measurement *n* times, then the probability of observing *m* detections is given by,

$$P(X = m; p, n) = p^m (1-p)^{n-m} \frac{n!}{m!\,(n-m)!}$$

3. The expected value of a binomial distribution is $\langle X \rangle = n\,p$, with a standard deviation of $\sigma = \sqrt{n\,p\,(1-p)}$.

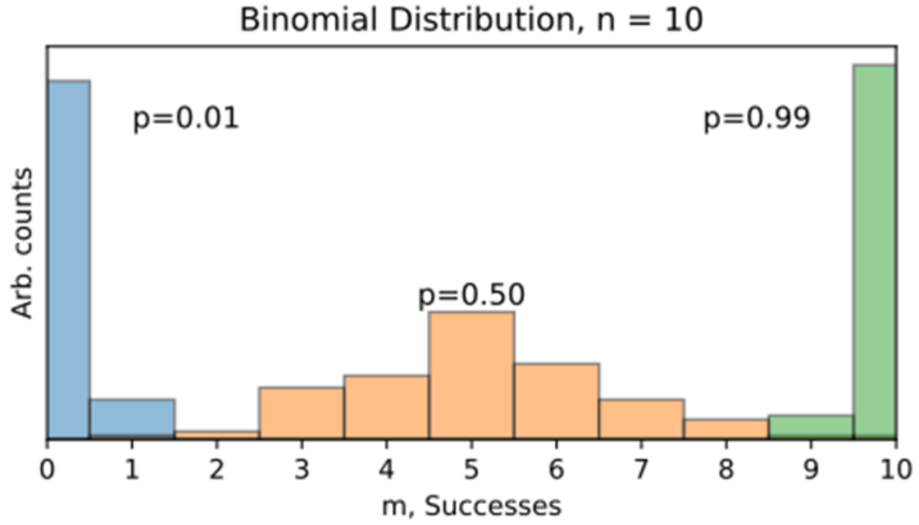


Figure B-2: Depiction of the variation in the number of observed successes, m, out of n = 10 attempts for various probabilities, p, repeated 100 times. Each single test against the 10 targets would yield a single value for the number of success, m. It is only by repeated measurements that the underlying probability distribution becomes apparent.

4. Consider a test lane with 10 targets, *n = 10*. Figure B-2 depicts the variation in the number of observed successes, *m*, for various probabilities, *p,* if we were to repeat that test many times. Each single test against the 10 targets would yield a single value for the number of success, *m*. It is only by repeated measurements that we build up the picture in Figure B-2.

5. For low probability of success, *p = 1%,* it is unsurprising that repeated trails of 10 targets would yield mostly 0 successes. Similarly, for a high *p = 99%* repeated trails would yield almost all successes. What is interesting is the wide variability in the potential outcomes when *p = 50%.*

6. Note, the experimenter does not usually know the value of *p.* In fact, it is most likely that *p* is the parameter of interest, and it is the observations of *m* successes in *n* tests that would lead to the *estimate* of *p,* $\hat{p} = m/n.$ If one were to run a single test of 10 targets, an observation of *m = 1* successes would suggest a low probability, but as shown in Figure B-2 it would also be consistent with *p = 50%.* The concept of *confidence* in the prediction of the true value of *p* given *m* successes in *n* tests will be discussed later as the binomial confidence interval.

B.2.4.4.2 Poisson

1. The Poisson distribution describes cases where a specific outcome occurs, but the number of trials is unknown. A metal detector searching a route may find a certain number of targets of interest, but it is meaningless to ask how much of the area it did not detect anything in. The Poisson distribution deals with the rate of events, and so is applicable to understanding the false alarm rate (FAR) of a military search sensor.

2. Let *X* be a Poisson random variable, denoted $X \sim Pois(\lambda)$, where λ is a rate, implying that an average λ events would be expected to occur in a search interval or search area. The probability mass function for a that search interval containing *k* events is,

$$P(X = k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

3. The expected value of the Poisson distribution $\langle X \rangle = \lambda$, with a standard deviation of $\sigma = \sqrt{\lambda}$.

4. Consider a test lane *2 m* × *100 m*, or $200\, m^2,$ in size. Figure B-3 depicts the number of expected false alarms that will be found if the false alarm rate is $\lambda = 1, 5, 10$ per $200\, m^2$. Each single test on this lane would yield a single value for the number of observed false alarms. It is only by repeated measurements that we build up the picture in Figure B-3.

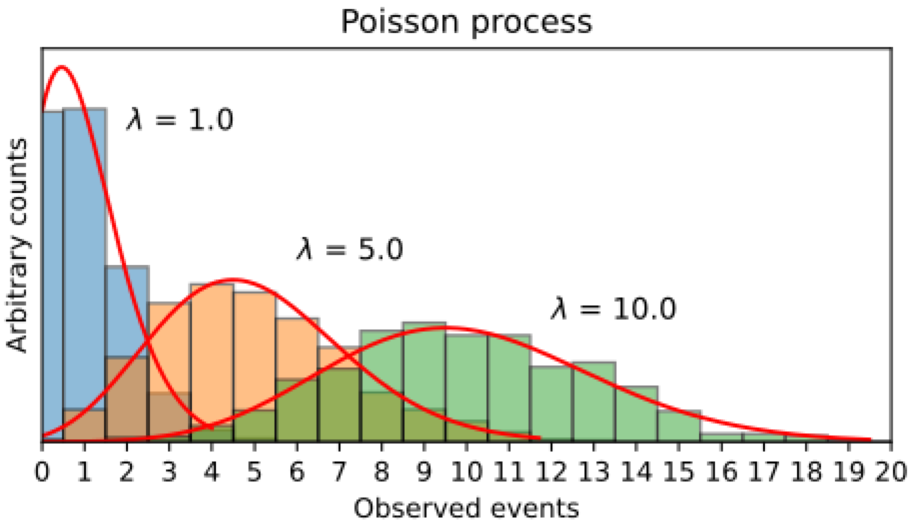


Figure B-3: Poisson mass density function for average rates $\lambda = 1, 5, 10$, in red. The histograms represent the expected outcomes for 100 repeated tests under these average rates. Included are the Normal approximations to the Poisson distributions (red solid lines). At low expected event rates the widths of the Normal approximation – which assume a symmetric data distribution – are seen to fail to represent the clearly asymmetric data.

5. For a ‘clean’ lane, that is a test lane that has been cleared of most, if not all, potential false alarms, the expected number of events is not unexpectedly low and varies minimally, as depicted by $\lambda = 1$ per $200\ m^2$. However, even with a modest expected rate of events, $\lambda = 5$ per $200\ m^2$, the expected variation in the observed number of events can be significant.

6. The challenge with the False Alarm Rate (FAR) when discussing the assessment and reporting of detection technologies used in military search is that it is inextricably linked to the specific trial site, local and temporal environment, and the systems under test. The FAR at two different sites, even for the same detection technology, are not meaningfully comparable unless the trial is specifically designed to assess that parameter. Differences in FAR for two systems tested at the same site is a reasonable question to ask, but even here one must recognize the observed false alarms may not be caused by the same phenomena; one detector may be seeing false alarms due to challenging soil conditions, while the other by ambient electromagnetic noise. Of all the statistical variables analyzed and reported in the testing of detection technologies, the FAR is by far the most difficult to quantify and leads to the most confusion and misinterpretation of results.

B.2.4.4.3 Normal

1. The Normal, or Gaussian, probability distribution of a random variable X, $X \sim \mathcal{N}(\mu, \sigma)$, is represented by a bell curve, symmetric about $x = \mu$ and a width controlled by a parameter σ, is described by,

$$\mathcal{N}(X = x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

2. The mean value of the Normal distribution is μ and the standard deviation is σ. Figure B-4 presents the Normal distribution with various parameters. The Standard Normal distribution is a special case obtained by setting $\mu = 0$, and $\sigma = 1$, denoted $\mathcal{N}(0,1)$,

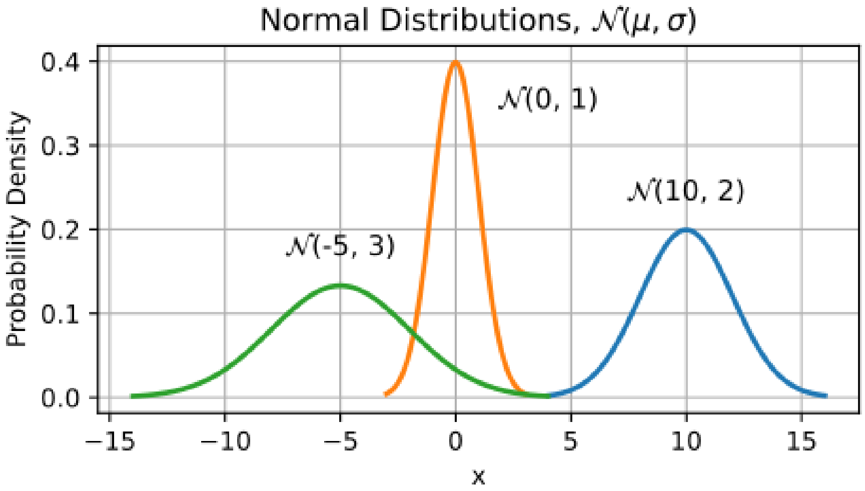


Figure B-4: Normal probability density function, as affected by changes in mean μ, and standard deviation σ.

3. Unlike the discrete binomial and Poisson distributions, which are concerned with the observation of a integer number of events (i.e. 'The probability of detecting 5 targets'), the Normal distribution is applicable to *continuous* variables and is interpreted as a *probability density*, which is only meaningful when considering ranges of the continuous variable, *x* (i.e. 'The probability of the range of detection falling between 25 – 35 cm').

4. As a *probability density,* the probability that a continuous variable observation *X* = *x* falls in the range [*a*,*b*] is calculated by,

$$P(a \leq X \leq b) = \int_a^b \mathcal{N}(x; \mu, \sigma)\, dx$$

5. As depicted in Figure B-5, this integral is the area under the curve from *a* to *b*. To make the Normal distribution consistent with the definition of a probability it is *normalized*, so that the integral over its range is 1,

$$\int_{-\infty}^{\infty} \mathcal{N}(x; \mu, \sigma)\, dx = 1.$$

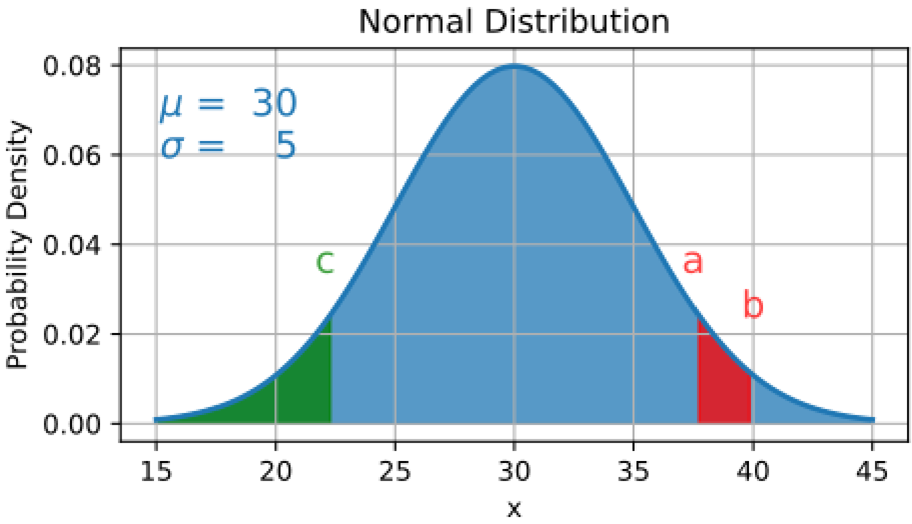


Figure B-5: In blue is plotted a Normal distribution with parameters μ= 30 and σ= 5. The integral from points [a,b] represents the probability of an observation X falling in that range. The Cumulative Integral evaluated at point c, F(c), represents the probability of an observation X ≤ c.

6. The *cumulative distribution function*, *F(X = x)*, is defined as the probability of a random observation being $\leq x$, defined as,

$$F(X = x) = \int_{-\infty}^{x} \mathcal{N}(y; \mu, \sigma)\, dy$$

7. Since the Normal distribution is normalized, then for point *c* in Figure B-5, $P(\, X \leq c\,) = \, F(c)$.

8. What does this mean? If one were to take repeated measurements of a process described by the probability distribution $X \sim \mathcal{N}(\mu, \sigma)$, then the fraction of observations with values in the range [*a*,*b*] would be described by $P(a \leq X \leq b)$. As mentioned earlier, it is unlikely that one knows in advance the exact probability distribution governing the measurements being made. Rather, it is most often the case that it will be the observations made that lead one infer the underlying probability distribution. Recall the previous discussion on histograms. If a histogram is viewed as being composed on many small intervals [*a*,*b*], then the fractional counts in a histogram bin approach the area under the curve in those intervals.

9. Consider Figure B-6. Plotted in red is a Normal distribution with parameters μ= 30 cm and σ= 5 cm, which recalls the earlier example of the maximum observed detection range of a detector to a target. From this probability distribution, 500 random events were generated and plotted in the histogram (blue). The statistical nature of randomly selecting events from an underlying distribution can be seen in the variations between histogram bins, but the general shape of the *parent* Normal distribution can be seen (red).

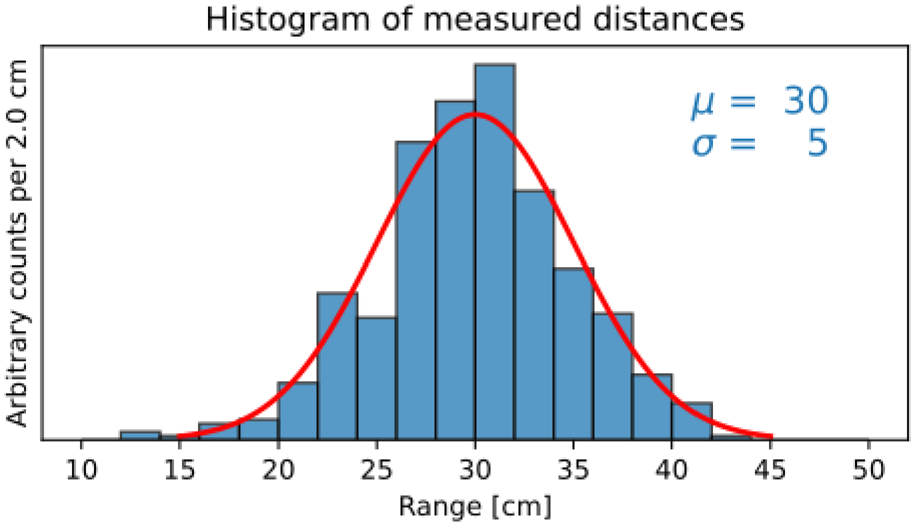


Figure B-6: In red is plotted a Normal distribution with parameters μ= 30 cm and σ= 5 cm. From this probability distribution 500 random events were generated and plotted in the histogram (blue). The statistical nature of randomly selecting events from an underlying distribution can be seen in the variations between histogram bins, but the general shape of the parent distribution (red) can be seen.

10. The Normal distribution is one of the most important and widely used distributions, not only because of its convenient analytic properties, but because of the *Central Limit Theorem*. The Central Limit Theorem implies that probabilistic and statistical methods that work for Normal distributions can be applicable to many problems involving other types of distributions, including the binomial and Poisson distributions.

B.2.4.5 Central Limit Theorem

1. The Central Limit Theorem (CLT) states that if you take the sum $X = \sum_{i=1}^{n} X_i$ of *n* independent random variables $\{X_i\}$, each taken from a distribution of mean μ_i and variance $V_i = \sigma_i^2$, then the distribution of X:

   a. has an expectation value of $\langle X \rangle = \sum_{i=1}^{n} \mu_i$;

   b. Variance $V(X) = \sum_{i=1}^{n} V_i = \sum_{i=1}^{n} \sigma_i^2$; and,

   c. becomes Normal as $n \to \infty$.

2. Why is this important? Because of the CLT, this holds true not just for Normal distributions, but others as well.

3. It can be shown that for a large number of trials, *n*, the binomial distribution tends to Normal with $\mu = np$ and a standard deviation of $\sigma = \sqrt{n\,p(1-p)}$, or $X \sim \mathcal{B}(p,n) \;\to X \sim \mathcal{N}(np, \sqrt{n\,p(1-p)})$. What constitutes large enough *n* is not well defined, although some guidelines have been developed. When p is not too

close to 0 or 1, then the Normal approximation is appropriate when [5], $n > 9\left(\frac{1-p}{p}\right)$, and, $n > 9\left(\frac{p}{1-p}\right)$. E.g., for a probability of detection of 70%, *n* should be greater than 21. For a probability of detection of 90%, *n* should be greater than 81.

4. Similarly, for a large number of trials the Poisson distribution tends to Normal with $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$, or $X \sim Pois(\lambda) \rightarrow X \sim \mathcal{N}(\lambda, \sqrt{\lambda})$.

5. Returning to the Normal distribution, suppose the same quantity is measured multiple times, perhaps the maximum range *x* of detection to a target, yielding a data series $\{x_i\}$. Each mutually exclusive sample is taken from the same distribution with the same mean $\mu_i = \mu$ and variance $V_i = \sigma^2$, thus, for the average value of $\bar{X} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$, the CLT reduces to its simple form and it can be shown [6],

   a. $\langle \bar{X} \rangle = \mu$

   b. $\langle V(\mathrm{X}) \rangle = \sigma_{\bar{X}}^2 = \sigma^2/\sqrt{n}$

6. This worth clarifying; the standard deviation, σ, represents the extent to which any single measurement x_i is expected to vary from the true mean μ, while $\sigma_{\bar{X}} = \sigma / \sqrt{n}$ is the standard deviation of the mean, and represents the extent to which the estimate $\bar{X}$ would vary from the true mean. This demonstrates that as more data samples are collected (larger *n*), the estimate of the mean remains the same, but the resolution, or error, of that estimate becomes smaller by a factor of $1/\sqrt{n}$ compared to a single measurement.

7. The effect of averaging is graphically depicted in Figure B-7. Here, samples of *n = 3, 10, 100,* and *1000*, were drawn from our example Normal distribution, $\mathcal{N}(30, 5)$. These experiments were repeated 100 times and the difference between the sample mean and distribution mean, $\bar{x} - \mu$, plotted. Not surprisingly, it can be seen that the mean of the average distribution, $\mu_{\bar{X}}$, approaches the distribution mean, $\mu_{\bar{X}} - \mu \rightarrow 0$. That is, increased data improves our estimate of the mean.

8. What is less obvious is that the standard deviation of the mean, $\sigma_{\bar{X}}$, can be less than the distribution mean, σ, and follows $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. In the example in Section B.2.3.1, for a test with *n = 50* measurements, the data set had mean $\bar{X}$ = 30.1 cm with standard deviation $\sigma =$ 4.3 cm, but the standard deviation on the estimate of the mean is only $\sigma_{\bar{X}} = 0.6$ cm. The *confidence* in the estimate of the mean is better than if it was based solely on the standard deviation of the data set.

9. The concept of confidence will be explored further below, but first *estimates* will be introduced before leaving the discussion on distributions.

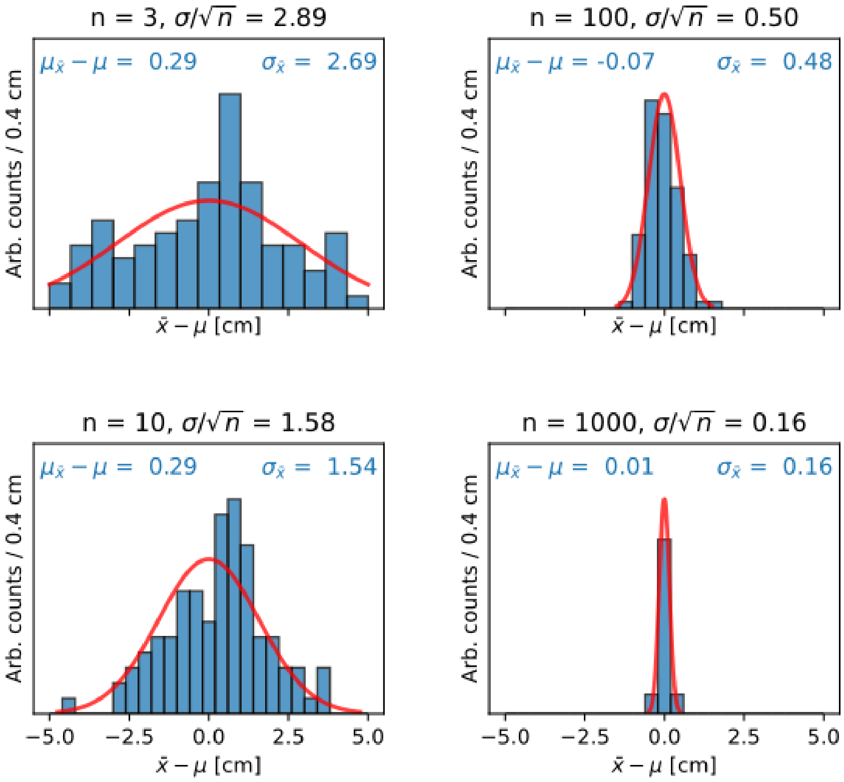


Figure B-7: Depiction of the reduction in the standard deviation of the estimate of the distribution mean with increased sample size. Samples of n = 3, 10, 100, and 1000, were drawn from a Normal distribution, $\mathcal{N}(30, 5)$. These experiments were repeated 100 times, and the difference between the sample mean and distribution mean, $\bar{X} - \mu$, plotted in blue. Overlaid in red are Normal distributions with $\mathcal{N}(\bar{X} - \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n})$.

B.2.4.6 Estimates

1. A common implicit assumption is that for a measurement being made there exists a true value, *x*, of that quantity that the measurement is trying to discern — i.e. there exists a true, repeatable range of detection for a metal detector against a specific metallic target, or true repeatable probability of detection against a particular target in a specific search scenario — and that each of the *n* observations in a sample $\{x_i\}$ represents an *estimate* of that true value but made inaccurate by experimental and random errors.

2. An *estimator,* denoted $\hat{x}$*,* is a procedure applied to a data sample which yields a numerical value, the *estimate*, of a parameter of the parent population based on observed data. While some estimators appear obvious — such as the arithmetic mean — there exist a wide variety of estimators defined for various distributions, but in all cases they are intended to be *consistent*, *non-biased*, and *efficient* [6]. These requirements may result in odd looking factors, such as a *(n-1)* where an *n* is expected, but their derivations are beyond the scope of this introduction.

3. The most relevant estimates of a data sample $\{x_i\}$ in regards to the analysis of military search detection technologies are: the mean, $\hat{\mu}$, the variance, $\hat{s}$, and the probability, $\hat{p}$, or rate, $\hat{\lambda}$, of the occurrence of a specific event. The following estimates (Table B-1) are calculable directed from the observed data sample $\{x_i\}$.

Table B-1: Principal estimators applicable to the assessment of Technologies for Military Search.

Estimators	Binomial	Poisson	Normal
Mean	$\hat{\mu} = n\hat{p}$	$\hat{\lambda} = k$	$\hat{\mu} = \bar{x}$
Standard Deviation	$\hat{s} = \sqrt{n\,\hat{p}(1-\hat{p})}$	$\hat{s} = \sqrt{\hat{\lambda}}$	$\hat{s} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$
Probability, or Rate	$\hat{p} = \frac{m}{n}$	$\hat{\lambda} = k$	$\hat{p}\,(X \leq c) = F(c;\, \hat{\mu}, \hat{s}\,)$

4. These estimates are based on the *observed* data, but are true only when the underlying distributions are correctly identified. For scenarios relevant to this discussion, this is typically achievable; binomial distribution is applied to tests with one of two outcomes, e.g. detect or not detect; Poisson distribution is applied to the False Alarm Rate; and, the Normal distribution is applied to tests of continuous variables, e.g. distances, or to tests with large populations that approach Normal due to the CLT.

B.2.5 CONFIDENCE

1. The Normal distribution allows one to better understand the meaning of the standard deviation, and how that is related to the reporting of statistical data.

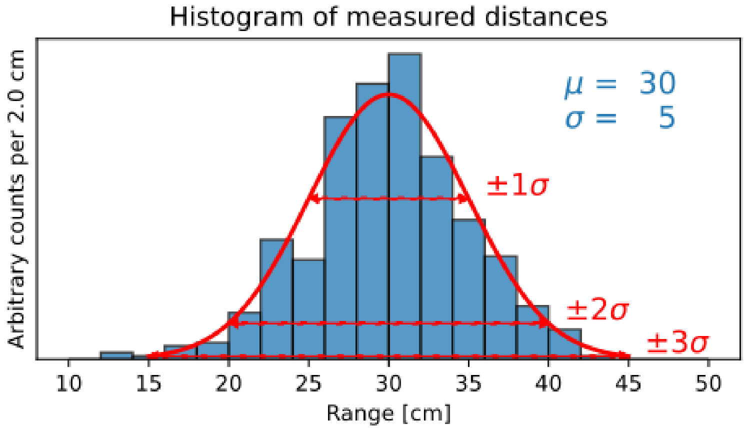


Figure B-8: A Normal distribution, μ= 30 cm and σ= 5 cm, overlaid with widths $\pm\sigma, \pm 2\sigma, and \pm 3\sigma$, representing 68 %, 95 % or 99.7 % of the total area under the curve.

2. Recalling Figure B-5, it was noted that the area under the curve over a range [*a*,*b*] represented an estimate of the probability of a random sample falling in that range. Figure B-8 depicts a Normal distribution overlaid with widths representing $\pm\sigma, \pm 2\sigma, \text{and} \pm 3\sigma$. It can be shown that the area under one standard deviation from the mean, $[\mu - \sigma, \mu + \sigma]$, is approximately 68% of the total area; two standard deviations $[\mu - 2\sigma, \mu + 2\sigma] = 95\%$; three standard deviations $[\mu - 3\sigma, \mu + 3\sigma] = 99.7\%$; and so on.

3. By interpreting these intervals as probabilities of occurrence, then if the random variable $X \sim \mathcal{N}(\mu, \sigma)$ is sampled on numerous occasions from the same population and a parameter estimate μ made on each occasion, the resulting estimates would bracket the true population parameter in approximately 68 %, 95 % or 99.7 % of the cases depending on whether the bracket was defined as $\pm\sigma, \pm 2\sigma, \pm 3\sigma$.

4. It is the variance, or standard deviation of the *observed* data that allows one to assign a degree of belief, or *confidence*, to a reported result. That is, an expression of the likelihood that the observation made represents the true underlying parameter. It's the difference between reporting definitively "The range of detection of Sensor A to Target B was 3.1 m," versus a description that relays the degree of belief in the result, which incorporates the scale of observed random variations that effected the measurement accuracy, "The range of detection of Sensor A to Target B was 3.1 ± 0.1 m," or even, if the experimental accuracy was poor, "The range of detection of Sensor A to Target B was 3 ± 2 m." This will be discussed more thoroughly in error analysis, but it is the cumulative distribution function of the Normal distribution that provides the underpinnings of the approach.

B.2.5.1 Normal Confidence Interval

1. As discussed above, if the random variable $X \sim \mathcal{N}(\mu, \sigma)$ is sampled on numerous occasions from the same population and the parameter estimate μ made on each occasion, the resulting estimates would bracket the true population parameter in approximately 68 %, 95 % or 99.7 % of the cases depending on whether the bracket was defined as $\pm\sigma, \pm 2\sigma, \text{or } \pm 3\sigma$. More generally, the question could be asked, “Given the observed data distribution, what is the range over which one could expect Y % of subsequent observations to fall?”

2. As previously depicted in Figure B-5, for any continuous probability distribution the probability density function is defined such that probability of a observation *x* falling in the range [*a*,*b*], $P(a \leq X \leq b)$, is represented by the area under the curve from points *a* to *b*,. The question then becomes, “What are the points [*a*,*b*] such that $P(a \leq X \leq b)$ = Y %?” These points are referred to as the *Z-critical factors.*

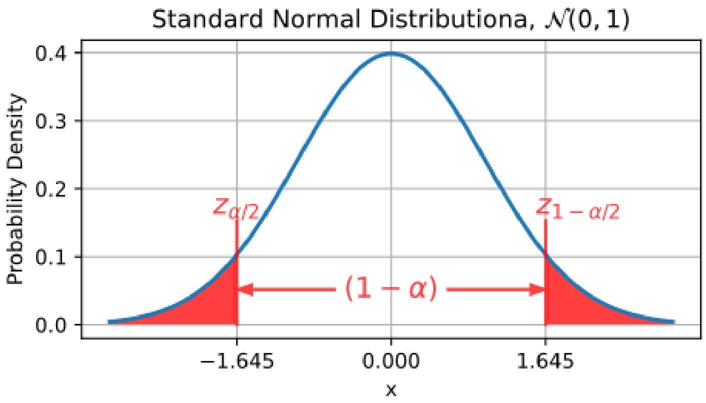


Figure B-9: Two-tailed critical factors.

3. Consider Figure B-9, which represents the Standard Normal Distribution, $\mathcal{N}(0,1)$. Following statistical convention, the *Z-critical factors* are defined in terms of their α value, which refers to the $100(1-\alpha/2)'$th percentile of the standard normal distribution. If one were to require, for example, $(1-\alpha) = 90\ \%$ of observations to fall within the defined interval, then $\alpha = 1 - 0.90 = 0.10$, represents the area outside that interval, shown in red. This total area is composed of two parts, or tails, each $\alpha/2$, which leads to the critical factors being labelled as $z_{\alpha/2}$ for the left tail, and $z_{1-\alpha/2}$ for the right tail. From the discussion on cumulative integrals in 5.4B.2.4.4.3, $z_{\alpha/2}$ is defined such that area to the left of $z_{\alpha/2}$ is $F(z_{\alpha/2}) = \alpha/2$, and $z_{1-\alpha/2}$ such that $1 - F(z_{1-\alpha/2}) = \alpha/2$. Strictly speaking, for $\mathcal{N}(0,1)$ $z_{\alpha/2}$ is negative, but for the symmetric $\mathcal{N}(0,1)$ $z_{\alpha/2}$ is often used as shorthand for its positive value $|z_{\alpha/2}| = z_{1-\alpha/2}$.

4. Z-critical factors for various distributions can be found in look-up tables, or calculated with readily available software. In this $\mathcal{N}(0,1)$ example, for $\alpha = 0.1$, $|z_{\alpha/2}| = z_{1-\alpha/2} = 1.64$. Again, due to the symmetry of the $\mathcal{N}(0,1)$, $|z_{\alpha/2}| =$

$z_{1-\alpha/2}$, allowing one to write more simply that range $\pm z_{\alpha/2} = \pm 1.64$ represents $(1-\alpha) = 90\ \%$ of the area. More generally, in the of case $\mathcal{N}(\mu, \sigma)$ this range is written as $\mu \pm z_{\alpha/2}\sigma$. In this form, the observations in Section B.2.5 above become more clear.

5. Critical values for common confidence intervals of 68%, 95%, and 99.7% are $z_{\alpha/2} = 0.99$, 1.96, and 2.97, or, approximately $\pm\sigma, \pm 2\sigma, \text{and} \pm 3\sigma$. For completeness, Table B-2 presents the Standard Normal Z-critical factors for common intervals, $68\%\ (\pm\sigma)$, 90%, $95\%\ (\pm 2\sigma)$, and $99.7\%\ (\pm 3\sigma)$:

Table B-2: Standard Normal Z-critical factors for common interval values.

	$(\mathbf{1-\alpha})$			
	0.683	**0.90**	**0.95**	**0.997**
$\boldsymbol{z_{1-\alpha/2}}$	1.00	1.64	1.96	2.97

6. Before leaving the discussion on the Normal confidence interval, it should be noted that this is generally held to be accurate for large *n*. For smaller *n*, say less than 10, it is recommended to calculate the Z-critical factors using the *Student T* distribution with degrees of freedom *n-1*. As with the Normal distribution, these can be found in look-up tables or calculable using readily available software.

7. The effort spent in defining Z-critical factors and understanding their genesis in the cumulative distribution function will become apparent below.

B.2.5.2 Binomial Confidence Interval

1. Aside from the mean and variance of an observed data sample $\{x_i\}$, the next most important quantity in regards to the analysis of military search detection technologies is the probability, *p*, of the occurrence of a specific event. The probability of detection is an extremely important value, with implications on technology acquisition, employment, and operational risk assessment, and so it should be well understood.

2. This is most commonly used in terms of detection; tests where either a target is detected, *d*, or not detected, 'not *d*'. In these cases, the binomial distribution $X \sim \mathcal{B}(p, n)$ is applicable, but it assumes one knows the probability in order to make predictions about expected events. In reality, however, it is most common that the investigator is observing events and trying to infer the probability.

3. If *m* detections were observed in *n* tests, then the probability of detection is estimated from observed data as $\hat{p} = \mathrm{m} \mathbin{/} \mathrm{n}$, but how should this be reported in order to express the belief in that estimate? Such an expression should in some way reflect the repeatability of the observed positive detections, which is related to the sample variance and the total number of trials conducted, *n,* so that an increase in the number of trials should increase our belief in the resulting

estimate. Further, it should express the confidence that value is somehow 'close' to the true value.

4. Unlike in the development of the Normal distribution Confidence Interval, the definition of the Binomial Confidence Interval is complicated by the fact that the binomial distribution is discrete. The exact solution is given by the following [7] [8].

5. The lower limit estimate of $\hat{p}$, $p_{LL}$is the value of *p* such that the sum of the probabilities of observing $> m$ events is less than a desired confidence, here denoted $\alpha/2$. That is, the fraction of the distribution tail *above* the observed successes *m* should be less than $\alpha/2$. This is equivalent to solving the following for p_{LL},

$$\sum_{k=m}^{n} p_{LL}^{k}\left(1-p_{LL}\right)^{n-k}\frac{n!}{k!\,(n-k)!}\leq\ \alpha/2\,.$$

6. Similarly, the upper limit p_{UL} is found by solving,

$$\sum_{k=0}^{m} p_{UL}^{k}(1-p_{UL})^{n-k}\frac{n!}{k!\,(n-k)!}\leq\ \alpha/2\,,$$

or, the fraction of the distribution tail *below* the observed successes *m*. The resulting range would be presented as $p_{LL} \leq p \leq p_{UL}$. This Clopper-Pearson formulation is inconvenient as it requires iterative computing algorithms to solve, thus more readily calculable mathematical approximations have been developed.

7. The most common method is given by the Wald method [2] [7],

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is the two tail *critical value* for a Standard Normal defined previously.

8. While the Wald formulation is useful for descriptive purposes, it is based on the Normal approximation of the binomial distribution and behaves poorly when *n* is too small, or where $\hat{p}$ is too close to an extreme, 0 or 1, as can be seen Figure B-10.

9. As trials of detectors for military search applications are often resource limited to smaller *n* than is desirable and the intent is to assess detectors with probabilities of detection that approach $p = 1$, the Wald formulation is a poor choice. There is a large and active body of research devoted to binomial confidence intervals in these extreme cases [7] [9], but these are quite mathematical and beyond the beyond the scope of this review. It is sufficient to present an alternative formulation and direct the reader to references if a more detailed explanation is required.

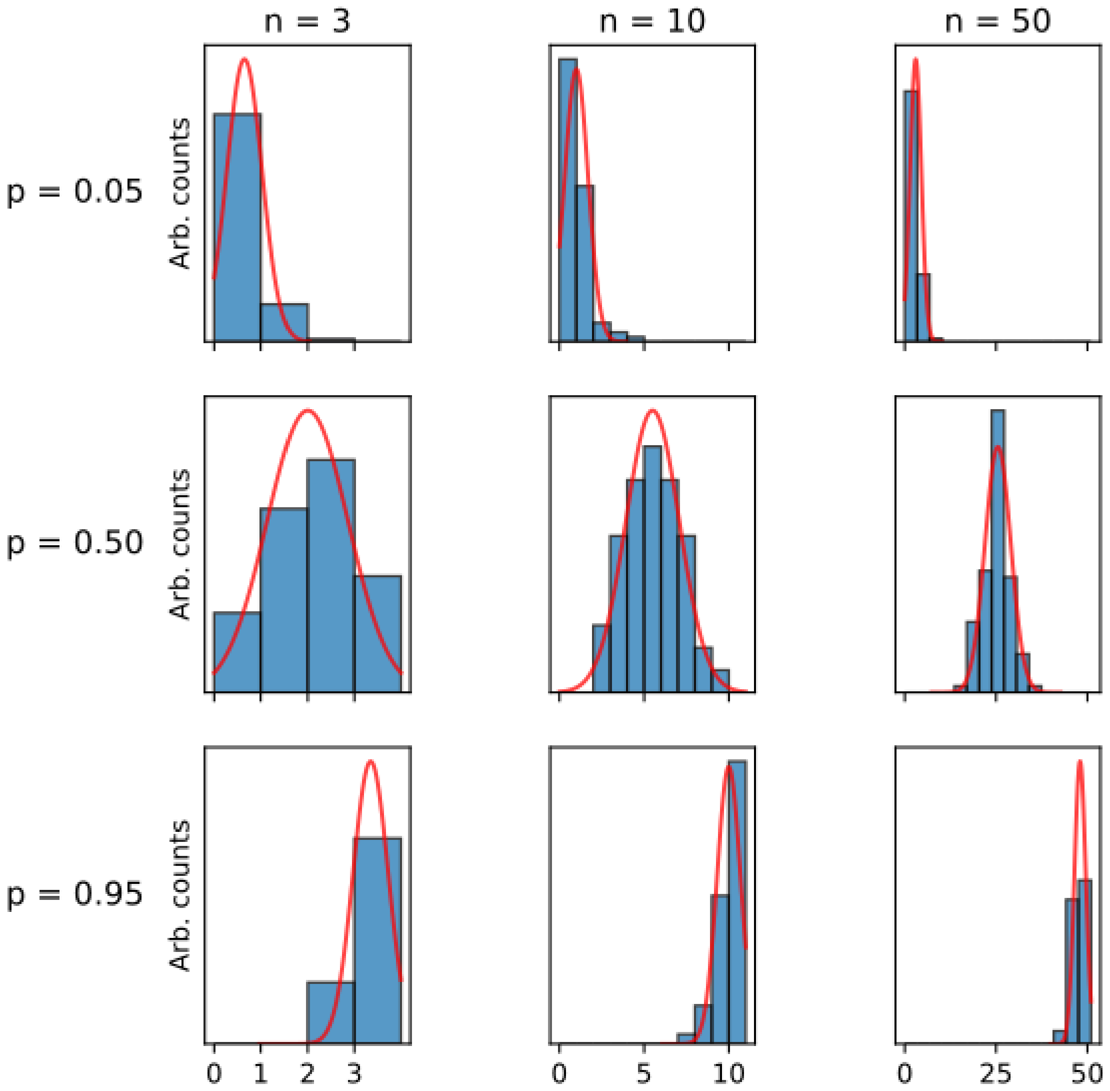


Figure B-10: Normal approximation (red) of the binomial distribution (blue). For moderately large n and p =0.5, the tails of the Normal distribution are well defined, thus the Z-critical factor in the Wald binomial confidence intervals are reasonable. As the n becomes small, or p approaches an extreme, 0 or 1, the Normal approximation becomes less applicable as the distributions become asymmetric and the tails of the Normal approximation become visibly effected.

10. More readily calculable and arguably better behaved than the Clopper-Pearson method, the Wilson method [7, 8] prescribes,

$$UL = \frac{\hat{p} + \frac{z_{1-\alpha/2}^2}{2n} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{z_{1-\alpha/2}^2}{n}}$$

$$LL = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

where $z_{1-\alpha/2}$ and $z_{\alpha/2}$ are the right and left tail critical values of the Standard Normal distribution.

11. In this spirit, a look-up table for the Modified Wilson approach is provided as being sufficiently justified for our purposes(Table B-3, Table B 4). From Table B-3, if an experiment observes 9 successes in 10 tests, the resulting 68% confidence interval would be reported as $0.7660 < \hat{p} = 0.9 < 0.9611$, or $\hat{p} = 0.9^{+\,0.0611}_{-\,0.1340}$.

12. The size of the binomial confidence interval bounds on small *n* may be surprising to readers familiar with reading reported detection rates from small *n* trials. 3 of 3 detections may appear promising, with an observed 100% probability of detection, but the true detection rate may be as low as 75% for a 68% ($\alpha = 0.317$) confidence interval (Table B-3), or even 44% for 95% ($\alpha = 0.05$) confidence interval (Table B-4). These windows do become smaller with increased *n*, but are still larger than 15% at *n* = 20, $\alpha = 0.05$.

Table B-3: Wilson Binomial Confidence Internal upper and lower limits for 68% coverage. For m successes in n tests, the upper limit UL is given by the first entry, followed by the lower limit LL, yielding bounds on the probability estimate, $LL < \hat{p} = \frac{m}{n} < UL$.

	m, Number of Successes																				
n, Number of Trials	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.5003	1																			
	0.0000	0.4997																			
2	0.3336	0.7888	1																		
	0.0000	0.2112	0.6664																		
3	0.2502	0.6145	0.8644	1																	
	0.0000	0.1356	0.3855	0.7498																	
4	0.2002	0.5002	0.7237	0.9001	1																
	0.0000	0.0999	0.2763	0.4998	0.7998																
5	0.1668	0.4209	0.6175	0.7841	0.9208	1															
	0.0000	0.0792	0.2159	0.3825	0.5791	0.8332															
6	0.1430	0.3631	0.5370	0.6891	0.8227	0.9344	1														
	0.0000	0.0656	0.1773	0.3109	0.4630	0.6369	0.8570														
7	0.1251	0.3192	0.4746	0.6128	0.7378	0.8495	0.9441	1													
	0.0000	0.0559	0.1505	0.2622	0.3872	0.5254	0.6808	0.8749													
8	0.1112	0.2846	0.4249	0.5510	0.6668	0.7732	0.8693	0.9512	1												
	0.0000	0.0488	0.1307	0.2268	0.3332	0.4490	0.5751	0.7154	0.8888												
9	0.1001	0.2568	0.3845	0.5001	0.6073	0.7073	0.8001	0.8844	0.9567	1											
	0.0000	0.0433	0.1156	0.1999	0.2927	0.3927	0.4999	0.6155	0.7432	0.8999											
10	0.0910	0.2340	0.3510	0.4576	0.5572	0.6508	0.7390	0.8212	0.8964	0.9611	1										
	0.0000	0.0389	0.1036	0.1788	0.2610	0.3492	0.4428	0.5424	0.6490	0.7660	0.9090										
11	0.0834	0.2148	0.3229	0.4217	0.5144	0.6022	0.6855	0.7644	0.8383	0.9062	0.9647	1									
	0.0000	0.0353	0.0938	0.1617	0.2356	0.3145	0.3978	0.4856	0.5783	0.6771	0.7852	0.9166									
12	0.0770	0.1986	0.2989	0.3910	0.4776	0.5601	0.6388	0.7139	0.7853	0.8524	0.9142	0.9677	1								
	0.0000	0.0323	0.0858	0.1476	0.2147	0.2861	0.3612	0.4399	0.5224	0.6090	0.7011	0.8014	0.9230								
13	0.0715	0.1846	0.2782	0.3643	0.4456	0.5232	0.5976	0.6691	0.7375	0.8027	0.8643	0.9210	0.9702	1							
	0.0000	0.0298	0.0790	0.1357	0.1973	0.2625	0.3309	0.4024	0.4768	0.5544	0.6357	0.7218	0.8154	0.9285							
14	0.0667	0.1725	0.2602	0.3411	0.4176	0.4908	0.5613	0.6292	0.6946	0.7575	0.8176	0.8744	0.9268	0.9724	1						
	0.0000	0.0276	0.0732	0.1256	0.1824	0.2425	0.3054	0.3708	0.4387	0.5092	0.5824	0.6589	0.7398	0.8275	0.9333						
15	0.0626	0.1618	0.2444	0.3206	0.3928	0.4621	0.5290	0.5936	0.6561	0.7165	0.7746	0.8303	0.8830	0.9318	0.9743	1					
	0.0000	0.0257	0.0682	0.1170	0.1697	0.2254	0.2835	0.3439	0.4064	0.4710	0.5379	0.6072	0.6794	0.7556	0.8382	0.9374					
16	0.0589	0.1524	0.2303	0.3024	0.3708	0.4366	0.5001	0.5616	0.6213	0.6793	0.7354	0.7895	0.8414	0.8906	0.9362	0.9759	1				
	0.0000	0.0241	0.0638	0.1094	0.1586	0.2105	0.2646	0.3207	0.3787	0.4384	0.4999	0.5634	0.6292	0.6976	0.7697	0.8476	0.9411				
17	0.0556	0.1440	0.2178	0.2862	0.3511	0.4136	0.4741	0.5328	0.5900	0.6455	0.6995	0.7519	0.8025	0.8511	0.8972	0.9400	0.9773	1			
	0.0000	0.0227	0.0600	0.1028	0.1489	0.1975	0.2481	0.3005	0.3545	0.4100	0.4672	0.5259	0.5864	0.6489	0.7138	0.7822	0.8560	0.9444			
18	0.0527	0.1365	0.2066	0.2716	0.3334	0.3930	0.4507	0.5068	0.5615	0.6148	0.6667	0.7173	0.7665	0.8140	0.8597	0.9031	0.9434	0.9786	1		
	0.0000	0.0214	0.0566	0.0969	0.1403	0.1860	0.2335	0.2827	0.3333	0.3852	0.4385	0.4932	0.5493	0.6070	0.6666	0.7284	0.7934	0.8635	0.9473		
19	0.0501	0.1298	0.1965	0.2584	0.3174	0.3742	0.4294	0.4831	0.5355	0.5867	0.6367	0.6855	0.7331	0.7794	0.8242	0.8673	0.9083	0.9464	0.9797	1	
	0.0000	0.0203	0.0536	0.0917	0.1327	0.1758	0.2206	0.2669	0.3145	0.3633	0.4133	0.4645	0.5169	0.5706	0.6258	0.6826	0.7416	0.8035	0.8702	0.9499	
20	0.0477	0.1237	0.1873	0.2464	0.3028	0.3572	0.4100	0.4615	0.5118	0.5610	0.6092	0.6563	0.7023	0.7472	0.7910	0.8334	0.8742	0.9130	0.9492	0.9807	1
	0.0000	0.0193	0.0508	0.0870	0.1258	0.1666	0.2090	0.2528	0.2977	0.3437	0.3908	0.4390	0.4882	0.5385	0.5900	0.6428	0.6972	0.7536	0.8127	0.8763	0.9523

Table B-4: Wilson Binomial Confidence Internal upper and lower limits for 95% coverage. For m successes in n tests, the upper limit UL is given by the first entry, followed by the lower limit LL, yielding bounds on the probability estimate, $LL < \ \widehat{p} = \frac{m}{n} < UL$.

		m, Number of Successes																				
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
n, Number of Trials	**1**	0.7935	1																			
		0.0000	0.2065																			
	2	0.6576	0.9055	1																		
		0.0000	0.0945	0.3424																		
	3	0.5615	0.7923	0.9385	1																	
		0.0000	0.0615	0.2077	0.4385																	
	4	0.4899	0.6994	0.8500	0.9544	1																
		0.0000	0.0456	0.1500	0.3006	0.5101																
	5	0.4345	0.6245	0.7693	0.8824	0.9638	1															
		0.0000	0.0362	0.1176	0.2307	0.3755	0.5655															
	6	0.3903	0.5635	0.7000	0.8124	0.9032	0.9699	1														
		0.0000	0.0301	0.0968	0.1876	0.3000	0.4365	0.6097														
	7	0.3543	0.5131	0.6411	0.7495	0.8418	0.9178	0.9743	1													
		0.0000	0.0257	0.0822	0.1582	0.2505	0.3589	0.4869	0.6457													
	8	0.3244	0.4709	0.5907	0.6943	0.7848	0.8632	0.9285	0.9776	1												
		0.0000	0.0224	0.0715	0.1368	0.2152	0.3057	0.4093	0.5291	0.6756												
	9	0.2991	0.4350	0.5474	0.6458	0.7333	0.8112	0.8794	0.9368	0.9801	1											
		0.0000	0.0199	0.0632	0.1206	0.1888	0.2667	0.3542	0.4526	0.5650	0.7009											
	10	0.2775	0.4042	0.5098	0.6032	0.6873	0.7634	0.8318	0.8922	0.9433	0.9821	1										
		0.0000	0.0179	0.0567	0.1078	0.1682	0.2366	0.3127	0.3968	0.4902	0.5958	0.7225										
	11	0.2588	0.3774	0.4770	0.5656	0.6462	0.7199	0.7873	0.8483	0.9025	0.9486	0.9838	1									
		0.0000	0.0162	0.0514	0.0975	0.1517	0.2127	0.2801	0.3538	0.4344	0.5230	0.6226	0.7412									
	12	0.2425	0.3539	0.4480	0.5323	0.6094	0.6805	0.7462	0.8067	0.8619	0.9111	0.9530	0.9851	1								
		0.0000	0.0149	0.0470	0.0889	0.1381	0.1933	0.2538	0.3195	0.3906	0.4677	0.5520	0.6461	0.7575								
	13	0.2281	0.3331	0.4223	0.5026	0.5763	0.6448	0.7086	0.7679	0.8229	0.8732	0.9182	0.9567	0.9863	1							
		0.0000	0.0137	0.0433	0.0818	0.1268	0.1771	0.2321	0.2914	0.3552	0.4237	0.4974	0.5777	0.6669	0.7719							
	14	0.2153	0.3147	0.3994	0.4759	0.5465	0.6124	0.6741	0.7320	0.7862	0.8366	0.8828	0.9243	0.9599	0.9873	1						
		0.0000	0.0127	0.0401	0.0757	0.1172	0.1634	0.2138	0.2680	0.3259	0.3876	0.4535	0.5241	0.6006	0.6853	0.7847						
	15	0.2039	0.2982	0.3788	0.4519	0.5195	0.5829	0.6425	0.6988	0.7519	0.8018	0.8482	0.8910	0.9295	0.9626	0.9881	1					
		0.0000	0.0119	0.0374	0.0705	0.1090	0.1518	0.1982	0.2481	0.3012	0.3575	0.4171	0.4805	0.5481	0.6212	0.7018	0.7961					
	16	0.1936	0.2833	0.3602	0.4301	0.4950	0.5560	0.6136	0.6682	0.7200	0.7690	0.8152	0.8584	0.8982	0.9341	0.9650	0.9889	1				
		0.0000	0.0111	0.0350	0.0659	0.1018	0.1416	0.1848	0.2310	0.2800	0.3318	0.3864	0.4440	0.5050	0.5699	0.6398	0.7167	0.8064				
	17	0.1843	0.2698	0.3434	0.4103	0.4726	0.5313	0.5870	0.6399	0.6904	0.7383	0.7839	0.8269	0.8672	0.9044	0.9381	0.9671	0.9895	1			
		0.0000	0.0105	0.0329	0.0619	0.0956	0.1328	0.1731	0.2161	0.2617	0.3096	0.3601	0.4130	0.4687	0.5274	0.5897	0.6566	0.7302	0.8157			
	18	0.1759	0.2576	0.3280	0.3922	0.4521	0.5087	0.5625	0.6138	0.6628	0.7097	0.7544	0.7969	0.8372	0.8750	0.9100	0.9416	0.9690	0.9901	1		
		0.0000	0.0099	0.0310	0.0584	0.0900	0.1250	0.1628	0.2031	0.2456	0.2903	0.3372	0.3862	0.4375	0.4913	0.5479	0.6078	0.6720	0.7424	0.8241		
	19	0.1682	0.2464	0.3139	0.3757	0.4333	0.4879	0.5399	0.5896	0.6372	0.6829	0.7267	0.7686	0.8085	0.8464	0.8819	0.9149	0.9448	0.9706	0.9906	1	
		0.0000	0.0094	0.0294	0.0552	0.0851	0.1181	0.1536	0.1915	0.2314	0.2733	0.3171	0.3628	0.4104	0.4601	0.5121	0.5667	0.6243	0.6861	0.7536	0.8318	
	20	0.1611	0.2361	0.3010	0.3604	0.4160	0.4687	0.5190	0.5671	0.6134	0.6579	0.7007	0.7418	0.7812	0.8188	0.8545	0.8881	0.9193	0.9476	0.9721	0.9911	1
		0.0000	0.0089	0.0279	0.0524	0.0807	0.1119	0.1455	0.1812	0.2188	0.2582	0.2993	0.3421	0.3866	0.4329	0.4810	0.5313	0.5840	0.6396	0.6990	0.7639	0.8389

13. Figure B-11 reproduces Figure B-2, with the Wilson 68% (α=0.317) confidence intervals. Recall that an investigator would only observe a single value of *m* success out of *n* trials, and from that would infer the underlying probability as $\hat{p} = m / n$. The presented intervals were calculated at specific values of observed successes, m = 0, 5, 10. Included are the Normal approximation to the binomial distribution (red solid lines). At the extremes, p = 0.01 and 0.99, the widths of the Normal approximation – which assume a symmetric data distribution – are seen to fail to represent the clearly asymmetric data, and resulting Wald intervals would underestimate the allowed range of the parameter *p*. The binomial confidence interval is better represented by the Wilson approach.

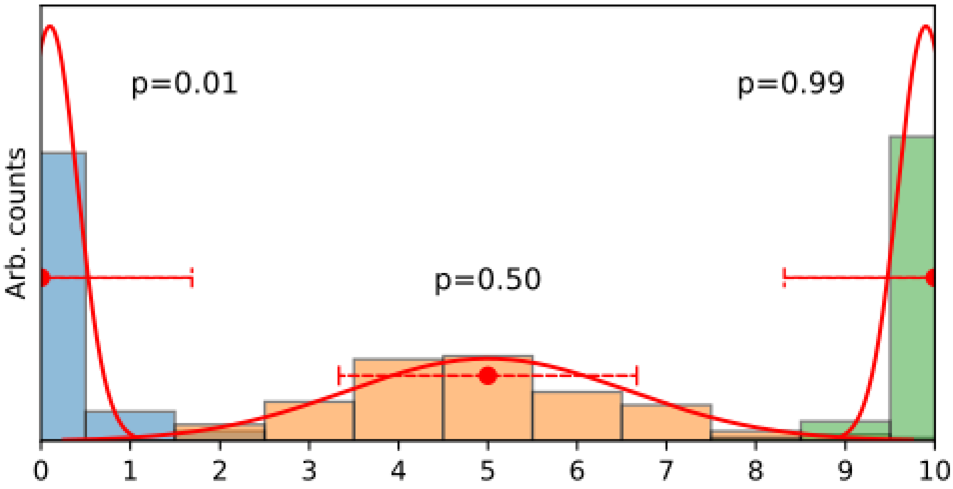


Figure B-11: Wilson 68% ($\alpha = 0.317$) confidence intervals (red dashed lines) overlaid on histograms of n = 10 binomial tests with various probabilities, p. These intervals were calculated at specific values of observed successes, m = 0, 5, 10. Included are the Normal approximations to the binomial distributions (red solid lines). At the extremes, p = 0.01 and 0.99, the widths of the Normal approximation – which assume a symmetric data distribution – are seen to fail to represent the clearly asymmetric data and would underestimate the allowed range of the parameter p. The binomial confidence interval for low n and extreme p is better represented by the asymmetric Wilson approach.

14. In regards to the alternative advanced methodologies, the differences between all the alternative with respect to the assessment and reporting of military search technologies do not appear significant. The most valuable outcome of this discussion will be if investigators and technical authorities begin to report confidence intervals in addition to the observed probabilities of detection. The fine details of these intervals, and their slight variations dependent on the specific interval definition chosen, would then only be a secondary consideration.

B.2.5.3 Poisson Confidence Interval

1. The definition of Poisson Confidence Intervals are similarly complicated by the fact that the Poisson distribution is also discrete.

2. Before continuing, it can be shown that the sum of $X \sim Pois(\lambda_i)$ processes is itself a Poisson process $X \sim Pois\left(\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)$, thus if one were to repeat *n* trials against the same target set in the same scenario, that is $\lambda_i = \lambda,$ then the expected number of events $\langle X \rangle = n\,\lambda$.

3. Suppose then that $\{k_i\}_{i=1}^{n}$ events were observed to occur in a series of *n* trials conducted over a search interval or search area. If the estimate of the rate is given by $\hat{\lambda} = \frac{1}{n}k = \frac{1}{n}\sum_{i=1}^{n}k_i$ then the exact form of the $(1-\alpha)$ Poisson confidence interval $\lambda_{LL} \le \lambda \le \lambda_{UL}$, can be defined by [11]: λ_{LL}, as the minimum value of the rate such that,

$$\sum_{i=k}^{\infty}\frac{e^{-\lambda_{LL}}\lambda_{LL}^{i}}{i!} \le \alpha/2\,;$$

and, λ_{UL}, as the maximum value of the rate such that,

$$\sum_{i=0}^{k}\frac{e^{-\lambda_{UL}}\lambda_{UL}^{i}}{i!} \le \alpha/2\,.$$

4. This interval can be exactly represented by [9],

$$\frac{1}{2}\mathcal{X}^2\left(\frac{\alpha}{2};2k\right) \le n\,\hat{\lambda} \le \frac{1}{2}\mathcal{X}^2\left(1-\frac{\alpha}{2};k+2\right)\,,$$

where $\mathcal{X}^2(p;n)$ is the lower tail area *p* of the Chi-squared distribution with *n* degrees of freedom. The interval for λ, $\lambda_{LL} \le \lambda \le \lambda_{UL},$ can then be calculated by scaling the previous result by *n*. This is readily calculable with modern software, but many argue that it is overly conservative in the resulting confidence interval estimate [12] [13] [14].

5. As with the binomial distribution, various readily-calculable approximations have been developed that also purport to address concerns such as the conservativeness, or not, of the resulting confidence interval but in reviewing these alternatives one needs to be mindful of their applicability in the low-λ region that most assessments of military search technologies will encounter; most operators require a low false alarm rate thus candidate search technologies tend to be tested in trials that produce a low number of false alarms.

6. For completeness, a readily calculable simplification has been shown to be quite accurate and applicable in the low-λ region, $\lambda \leq 5$ [10]. For a given observation, $\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i$, drawn from *n* identical $X \sim Pois(\lambda)$, the *Scores* $(1 - \alpha)$ confidence interval for λ is given by [11],

$$\bar{k} + \frac{z_{\alpha/2}^2}{2n} - \frac{z_{\alpha/2}}{\sqrt{4n}}\sqrt{\frac{4\bar{k} + z_{\alpha/2}^2}{n}} \leq \lambda \leq \bar{k} + \frac{z_{\alpha/2}^2}{2n} + \frac{z_{\alpha/2}}{\sqrt{4n}}\sqrt{\frac{\bar{4}k + z_{\alpha/2}^2}{n}} \;,$$

where $z_{\alpha/2}$ is the upper tail Z-critical value of the Standard Normal distribution.

7. Figure B-12 reproduces Figure B-3, using the $\mathcal{X}^2$ approach to calculate $(1 - \alpha) = 68\%$ confidence intervals. Recall that an investigator would only observe a single number of *k* alarms per trial, and from that would infer the underlying probability as $\hat{\lambda} = k$. The presented intervals were calculated at specific values of observed alarms, *k* = 1, 5, 10. Included are the Normal approximation to the Poisson distribution, $\mathcal{N}(\lambda, \sqrt{\lambda})$ (red solid lines). At low λ the widths of the Normal approximation – which assume a symmetric data distribution – are seen to fail to represent the clearly asymmetric data.

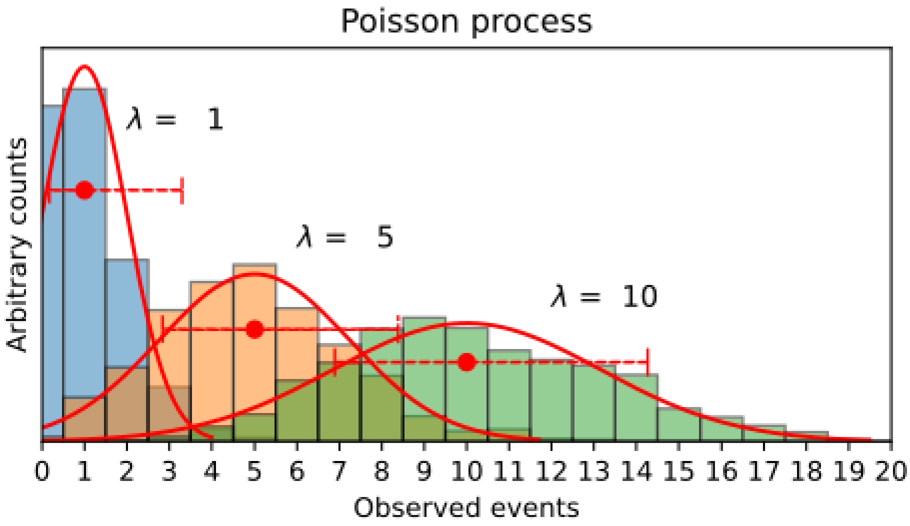


Figure B-12: Asymmetric 68% ($\alpha = 0.317$) Scores Poisson confidence intervals (red dashed lines) overlaid on histograms generated from Poisson processes with rates $\lambda = 1, 5, and\ 10$. Included are the Normal approximations to the Poisson distributions (red solid lines). At low expected event rates the widths of the Normal approximation – which assume a symmetric data distribution – are seen to fail to represent the clearly asymmetric data and would underestimate the allowed range of the rate parameter λ.

For completeness, look-up tables for both the $\mathcal{X}^2$ and *Scores* methods are provided in Table B-5 and

8. Table B-6. As an example, suppose a total of *k = 20* events were observed in a repeated series of *n* = 3 trials, yielding $\hat{\lambda} = \frac{k}{n} = 6.67$ events per search. From Table B-5, the $\mathcal{X}^2$ range for *k* = *20* at $(1-\alpha) = 0.68$ is

$$5.587 \le 3\,\lambda \le \; 25.518\,,$$

from which,

$$\lambda_{LL} = \lambda \; - \frac{\mathrm{n}\,\lambda - \,15.587}{n} \;\; = \; 5.196\,,$$

$$\lambda_{UL} = \lambda + \frac{(25.518 - \;\mathrm{n}\,\lambda\,)}{n} \;\; = \; 8.506 \;\;,$$

for *n* = 3, or, $5.196 \le \; \hat{\lambda} = 6.667 \le 8.509$. Note, the interval resulting from the sum of *n = 3* trials is less than the interval of a single trial with $k = 7 \;\cong\; 6.667$, which reinforces the value of obtaining extra data when possible.

Table B-5: Exact (χ^2) Poisson Confidence Intervals for k observed events at indicated (1-α) confidence levels.

	Chi-Squared					
	0.68		**0.90**		**0.95**	
k	**LL**	**UL**	**LL**	**UL**	**LL**	**UL**
0	0.000	1.833	0.000	2.996	0.000	3.689
1	0.174	3.289	0.051	4.744	0.025	5.572
2	0.712	4.625	0.355	6.296	0.242	7.225
3	1.373	5.904	0.818	7.754	0.619	8.767
4	2.093	7.147	1.366	9.154	1.090	10.242
5	2.849	8.365	1.970	10.513	1.623	11.668
6	3.630	9.566	2.613	11.842	2.202	13.059
7	4.429	10.751	3.285	13.148	2.814	14.423
8	5.243	11.925	3.981	14.435	3.454	15.763
9	6.069	13.089	4.695	15.705	4.115	17.085
10	6.905	14.245	5.425	16.962	4.795	18.390
11	7.749	15.394	6.169	18.208	5.491	19.682
12	8.600	16.536	6.924	19.443	6.201	20.962
13	9.457	17.673	7.690	20.669	6.922	22.230
14	10.320	18.805	8.464	21.886	7.654	23.490
15	11.188	19.933	9.246	23.097	8.395	24.740
16	12.061	21.057	10.036	24.301	9.145	25.983
17	12.937	22.177	10.832	25.499	9.903	27.219
18	13.817	23.293	11.634	26.692	10.668	28.448
19	14.700	24.407	12.442	27.879	11.439	29.671
20	15.587	25.518	13.255	29.062	12.217	30.888

Table B-6: Scores Poisson Confidence Intervals for k observed events at indicated (1-α) confidence levels.

	Scores					
	0.68		**0.90**		**0.95**	
k	**LL**	**UL**	**LL**	**UL**	**LL**	**UL**
0	0.000	0.989	0.000	2.706	0.000	3.841
1	0.384	2.605	0.223	4.482	0.177	5.665
2	1.004	3.985	0.662	6.044	0.548	7.293
3	1.702	5.286	1.199	7.507	1.020	8.821
4	2.445	6.544	1.796	8.910	1.556	10.286
5	3.216	7.772	2.434	10.272	2.136	11.706
6	4.009	8.980	3.103	11.603	2.750	13.092
7	4.817	10.172	3.795	12.910	3.391	14.451
8	5.639	11.350	4.508	14.198	4.054	15.788
9	6.470	12.519	5.236	15.469	4.735	17.106
10	7.311	13.678	5.978	16.727	5.432	18.409
11	8.159	14.830	6.732	17.973	6.142	19.699
12	9.014	15.975	7.496	19.209	6.865	20.977
13	9.875	17.114	8.270	20.436	7.598	22.244
14	10.741	18.248	9.051	21.654	8.340	23.502
15	11.611	19.378	9.840	22.865	9.091	24.751
16	12.486	20.503	10.636	24.070	9.849	25.992
17	13.365	21.624	11.437	25.268	10.614	27.227
18	14.246	22.742	12.244	26.461	11.386	28.455
19	15.132	23.857	13.057	27.649	12.164	29.677
20	16.020	24.969	13.873	28.832	12.948	30.894

9. As alluded to throughout the examples used in this section, the value of understanding and calculating confidence intervals from observed data is in the error analysis and reporting of experimental results.

B.2.6 ERROR ANALYSIS

1. Error analysis is the study and evaluation of uncertainty in measurement [12].

2. The importance of error analysis, and reporting of error, should now be clear. Not only does it inform the investigator as to the degree of belief in their results — especially important when informing technology acquisition and employment decisions — but as well the process of conducting error analysis prompts the investigator to thoroughly consider and report their work as clearly as possible, with the benefit of providing the reader the ability to interpret for themselves the value of their observations and recommendations.

3. Experimental error is unavoidable, but through thoughtful consideration and experimental design it can be minimized, and to some extent, quantified. Human-error in reading instrumentation, the finite resolution of measurement

apparatus, unaccounted-for variability in temperature during a long trial, and voltage fluctuations on a sensor components are all examples of the processes that contribute to inaccuracies in the collected data set, which are accumulated in the variance in the data set. Some errors are termed *statistical*, and others *systematic*.

B.2.6.1 Statistical Error

1. Assuming that an underlying physical parameter of interest, *x*, has a true value, µ, then it is supposed that the average of multiple measurements $\bar{x}$ will approach µ as the number of samples *n* becomes large.

2. Even when all experimental conditions are held constant, there will be irreducible variability in repeated measurements due to the accumulated effects of a variety of possible sources of small, uncorrelated errors. These sources of error will tend to combine randomly to produce measurements that vary about the true mean.

3. If *uncorrelated* (that is, not causally related), and random, then all these sources of error contribute to the observed variance in the collected data set, denoted σ_x^2, where the subscript *x* indicates it is the variance of the single measurements, and denotes the uncertainty of any one measurement x_i compared to the true value, µ.

4. It is intuitive that multiple samplings should improve our knowledge of the true parameter, as well as reduce the effects of random errors, which is what the Central Limit Theorem tells us (section B.2.4.5). This is reflected in the fact that the mean, $\bar{x}$, provides a better estimate of the true mean than does any single measurement, and how the standard deviation of the mean, $\bar{x}$, is given by $\sigma_{\bar{x}} = \sigma_x/\sqrt{n} \leq \sigma_x$.

5. To improve the confidence in a measurement, therefore, one can both improve the experimental technique in order to minimize σ_x, as well as increase the number of tests in order to calculate an average value with a standard deviation decreased by a factor of $1/\sqrt{n}$.

6. It was discussed previously that ± 1 standard deviation about the mean of the Normal distribution represents a 68 % confidence interval, so reporting the average of the measurements $\{x_i\}$ as $\bar{x} \pm \sigma_{\bar{x}}$, is stating that if the trial was to be conducted numerous times it is expected that the observed mean $\bar{x}$ would fall within $\pm\sigma_{\bar{x}}$ of the true mean, µ, 68 % of the time.

7. Given the observed data set collected, this simple expression captures both the best estimate of the true mean, $\bar{x}$, as well as an expression of confidence in that value based on the observed standard deviation; a small/large σ implying more/less confidence that the observed $\overline{x}$ accurately represents the true value parameter being measured, and thus relays to the reader some measure of appropriate belief in the results.

8. This random, or *statistical* error as it is known, is directly observed in the data as variance and can be reduced through increased measurement, unlike *systematic* error.

B.2.6.2 Systematic Error

1. Systematic errors are those sources of uncertainty that cannot be revealed by multiple measurements. One common cause of such error is calibration error. If there is a calibration error in an instrument — perhaps caused by a temperature dependent voltage gain that remains uniform over a single trial, but changes daily as the temperature fluctuates — that error will be present with the same value in all subsequent measurements. Multiple measurements with the same instrument settings will not average that error contribution to zero because is not random.

2. For errors such as this, it is important to thoroughly identify and consider possible sources of experimental uncertainty, and quantify them separately through dedicated tests. These tests would be intended to ascertain the scale of potential contributions of uncertainty to an experimental results due to variables not otherwise controlled of measured during the experiment.

3. Consider again the example presented in section B.2.3.1, where the repeated measurements of the maximum detection distance of a detection technology to a specific target was found have a mean $\bar{x} = 30.0$ cm and a standard deviation σ = 4.3 cm. If that detection technology included the setting of an alarm threshold once per trial series, then any sensitivity of the measured detection distance to that setting would be present consistently, not randomly, throughout those measurements. That is, if a particular threshold setting off-set the distance measurements by 1 cm, then all measurements would be offset by this same amount. As the contribution is not random, it will not be captured by the observed data variation.

4. If one knew what this off-set was prior to the measurement series then it could be accounted for in the results, however, what if the impact of the threshold on any particular trial was unknown? One could first investigate the sensitivity of the measured detection distance to threshold setting by conducting a separate, detailed test. If this test showed that the variation in repeated identical measurements, differing only by the repeated re-setting of the threshold, yielded a standard deviation of 0.1 cm, then that would the *systematic* error that should be combined with the *statistical* error of any subsequent trial series, as discussed next.

B.2.6.3 Error Combination

1. If σ_{stat} and σ_{sys} represent the statistical and systematic uncertainties determined from the preceding assessments, then the measured value of parameter *x* can be reported as

$$x \pm \sigma_{stat} \pm \sigma_{sys} ,$$

representing the most detailed assessment of parameter x obtained from the investigation.

2. By separating out the sources of error in this fashion, the reader is presented the opportunity to interpret the result within the relative contributions of the sources of error, which may lead to decisions to accept as reported, repeat, contribute to, or improve upon an experimental investigation. However, this format has a tendency to mislead the non-expert reader into believing the true confidence in the result is smaller than it is, because the two separate errors will most often be smaller than their combined effects.

3. Alternatively, the argument can be made that since statistical and systematic sources of error are typically uncorrelated, they can be combined in quadrature, as

$$\sigma_{total}^2 = \sigma_{stat}^2 + \sigma_{sys}^2,$$

$$\sigma_{total} = \sqrt{\sigma_{stat}^2 + \sigma_{sys}^2}\ .$$

4. The resulting best estimate of parameter *x* would be reported as $x \pm \sigma_{stat} \pm \sigma_{sys}$, or more conveniently but less accurately, $x \pm \sigma_{total}$.

5. This cannot be rigorously justified, nor can one claim the resulting confidence interval is accurately defined, but one can see that σ_{total} will always be greater than or equal to the contributions of statistical and systematic errors and so gives a reasonable indication of the overall confidence in the experimental result. For tests of detection technologies used in military search, it is the relative size of the reported uncertainty that is usually of interest, not accurate confidence interval bounds, thus σ_{total} is often more convenient.

B.2.6.4 Error Reporting

1. It has been suggested that an estimate of parameter *x* would be reported in the form $x \pm \sigma$, but what magnitude of σ should be used? In previous discussion on confidence intervals it was noted that the common options for symmetric distributions are ±σ ($(1-\alpha) = 68\%$), ±1.64σ (90%), ±2σ (95%), and ±3σ (99.7%). For nonsymmetrical distributions, the $(1-\alpha)$ choice would be the same but the error would be reported as $x^{+\,\Delta_{UL}}_{-\,\Delta_{LL}}$, when $\Delta_{UL} \neq \Delta_{LL}$.

2. The appropriate selection of $(1-\alpha)$ is left to the investigator. It is suggested that if the investigator has no specific requirement to express confidence levels to a high value of $(1-\alpha)$, such as “The probability of detection is greater than 98% at $(1-\alpha) = 0.95$”, then ±σ, or $(1-\alpha) = 68\%$, is sufficient to relay the magnitude of the experimental error.

B.2.7 EXPERIMENTAL INTERPRETATION

3. With the statistical tools now at hand, the investigator is able to:

a. report the observed mean, sample standard deviation, and error of the mean obtained from a measurements of continuous random variable, often described as Normally distributed $X \sim \mathcal{N}(\hat{\mu}, \hat{s})$, representing a physical parameter that yields a data set $\{x_i\}_{i=1}^{n}$, yielding, $\bar{x} \pm \sigma/\sqrt{n}$;

b. report on the observed probability and binomial confidence intervals of a binomially distributed random variable $X \sim \mathcal{B}(\hat{p}, n)$ observing *m* successes in *n* tests, $p_{LL} < \hat{\mathrm{p}} = \frac{m}{n} < p_{UL}$; and,

c. report on the observed false alarm rate and Poisson confidence intervals of a Poisson distributed random variable $X \sim Pois(\hat{\lambda})$, observing *k* background events in *n* tests, $\lambda_{LL} < \hat{\lambda} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} k_i < \lambda_{UL}$.

4. This knowledge will enable the technical authority to share and understand the significance of their observations while expressing the limits of the inferences that should be drawn from them.

5. The investigator is next expected to interpret these results in order to influence capability development decisions. In regards to the assessment of continuous parameters, such as the maximum detection range to a target, analysis beyond the conclusions reached above delves into *statistical significance* and *sensitivity*, which are beyond the scope of this introduction. The detection performance of a system against a specified target set, however, leads one to explore the often correlated effects on probability of detection and false alarm rates that can be investigated further.

B.2.7.1 Truth Table

1. For any detector-variable-target test scenario, the resulting observations can be categorized according to a truth table as defined in Table B-7:

Table B-7: Truth Table

	Target Present	**Non-Target**
Detection	True positive $(1-\alpha)$	False positive (β)
Rejection	False negative (α)	True negative $(1-\beta)$

2. Where,

a. True Positive – The detector functioned as desired by detecting the intended target. This contributes to the detector’s Probability of Detection (PD) evaluation, where PD is defined as the number of detected targets divided by the total number of targets.

b. False Negative – A target that failed to be detected. This represents a potentially dangerous situation in operations.

c. False Positive – Detector alarms against a non-target object. Contributes to the False Alarm Rate (FAR), which is a standard reporting metric for military search equipment.

d. True Negative – Common operating condition.

3. Figure B-13 will further clarify these definitions.

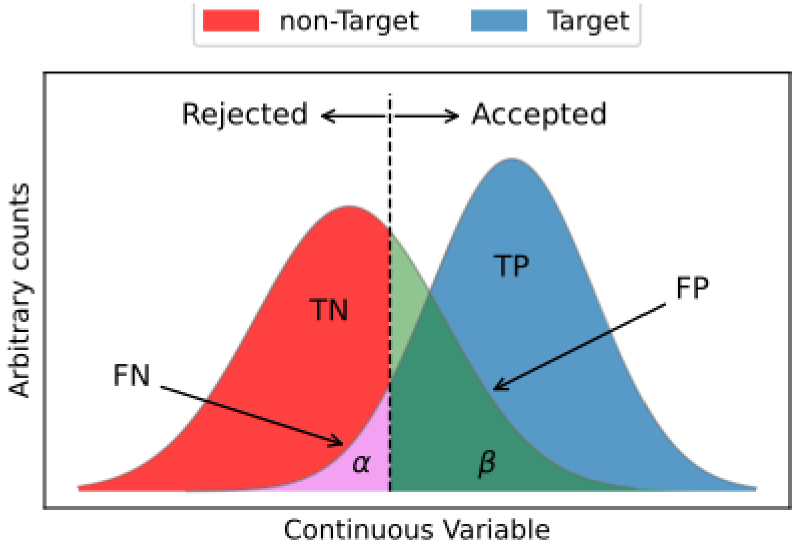


Figure B-13: The measured distributions of a continuous physical parameter from both targets and non-target objects are shown, such as the induced current reading on a metal detector coil, or the voltage measured on a radar receiver. By applying a decision threshold value to the physical parameter — with objects being either detected, if the measurement is above that threshold, or rejected, if that measurement is below that threshold — then the distinction between target and non-target can be developed.

4. As depicted in Figure B-13 this target/non-target differentiation is often imprecise, and almost any selection of the decision threshold value will come at the risk of either missing targets or accepting non-targets. In this scenario, the target events accepted as being above the threshold cut are True Positives (TP), and those being rejected are False Negatives. Recalling the discussion in Section B.2.4.4.3, the area of the target distribution below the threshold cut represents the cumulative probability of the target set being classified as False Negatives (FN).

5. The effect of the decision threshold on non-target objects is similarly defined. Those rejected by the threshold are True Negatives (TN), and those accepted are False Positives (FP), also known an False Alarms.

6. One possible decision criterion that the investigator could set when assessing search technologies is that the probability of observing a FN be less than α, implying that the probability of observing a TP, or PD, is $(1-\alpha)$, which is in fact the origin of the notation used previously.

7. A second consideration for the investigator may be the probability of observing a FP. This is depicted as the upper tail cumulative probability, β, so that the

investigator may require the probability of FP be less than a specified β, implying that the probability of observing a TN is (1-β).

8. The rates of True Negatives and False Positives when assessing a specific detector's performance is an expression of *statistical significance* [1], and is an important consideration in the design of experimental trials, particularly when identifying the required sample size *n* necessary to achieve the desired clarity on detector performance. Perhaps more importantly, while not frequently discussed specifically in assessments of military search technologies, statistical significance is often implied when comparing the performance of different systems; reports making direct comparisons of the TP (PD) and FP (FAR) of two systems but without strict analysis of whether these differences are truly statistically significant as opposed to merely consistent with expected statistical fluctuation. Further development of statistical significance is beyond the scope of this introduction, but here again it is noted that the larger the data sample *n* the less likely incorrect inferences will be made.

9. On a related note, for those detection technologies concerned with detecting quantities of target – such as ionizing radiation or trace vapour detectors – the concepts developed above also provide the fundamentals for defining the *sensitivity*, *detection limit*, or *lower limit of detection* [17] [18] of those detection technologies. However such analyses can be quite detailed and as those technologies represent specialized military search requirements it will not be pursued here.

10. Lastly, it should be emphasized that False Negative events represent a potentially dangerous situation in operations – targets missed – thus their minimization is often prioritized in search technology requirements. However, this often comes at the expense of resources, for to increase the PD often implies a concurrent increase in FP, and thus a higher FAR. With increased FAR more mission resources (time, equipment, personnel) are used to investigate false targets, and operator confidence in the systems may be reduced. In operational scenarios where false alarms are rare, the overall FAR will be low and so a higher PD can likely be obtained without significantly impaction operational tempo. In scenarios where false alarms are high, reduced PD may be acceptable to maintain the overall search rate, or the use of additional complimentary sensors that reduce the FAR through *sensor fusion* may be considered.

B.2.7.2 False Alarm Rate

1. The False Alarm Rate deserves particular attention in this review, as it is often miscommunicated or misinterpreted in assessments of military search technologies.

2. Fundamentally, a detection is declared when an object capable of being detected — be it by design, or material construction — is detected. That is, a detection is defined by the physical operating principles of the detection technology itself, not by the intent of the detector's use. For example,

uninteresting battlefield metallic clutter would be detected by a military metal detection sensor, although the principal operational intent of that system would be to locate explosive hazards. In a military context, such battlefield metallic clutter would commonly be referred to as a false alarm.

3. This is a common source of confusion when assessing technologies that detect secondary properties of threats, such as electromagnetic induction sensors (EMI, or metal detectors) when used as “landmine” detectors. EMI detectors can have a very low *technology false alarm rate* when considered as metal detectors – they can find very small quantities of metal with few false alarms – but can have a very high *application false alarm rate* as landmine detectors when operating in environments strewn with metal clutter, as one would find on a battlefield.

4. This only highlights the need for the investigator to be very clear in their experimental intent and reporting, so as to avoid such miscommunication.

B.2.7.3 Receiver Operating Characteristics

1. The acceptable balance between detection of targets, rejection of non-targets, PD, and FAR is not universally defined, rather it is a decision made by the investigator based on the operational requirements of the search equipment and the environments they will be used in. However, one way to investigate the trade-off between these various pressures is the Receiver Operating Characteristics (ROC).

2. As introduced above, Figure B-14(a) presents the measured distributions of a continuous physical parameter from both targets and non-target objects. A decision *threshold* is applied to the physical parameter such that objects are either detected, if the measurement is above that threshold, or rejected, if that measurement is below that threshold. Depicted in Figure B-14(a), this differentiation is often inefficient and almost any selection of the threshold value, with examples shown as $t_1, t_2, \text{and } t_3$, will come at the risk of either missing targets or accepting non-targets.

3. This can be visualized more clearly by considering the *1 - Cumulative* distribution, depicted in Figure B-14(b). The *1 - Cumulative* distribution presents the percentage of the parent distribution that is greater than the threshold cut at that value. That is, it is related to the number of targets that would be detected, and the number of non-target objects accepted, at that threshold. A cut at t_1 would detect almost all target objects, with the blue *1 - Cumulative* distribution almost 100%, but would admit a possibly unacceptable percentage of non-target items, here almost 60%. A cut at t_3 on the other hand would significantly reduce the probability of non-target objects being accepted, but at a significant cost to the probability of detection for a target object, here falling to 50%.

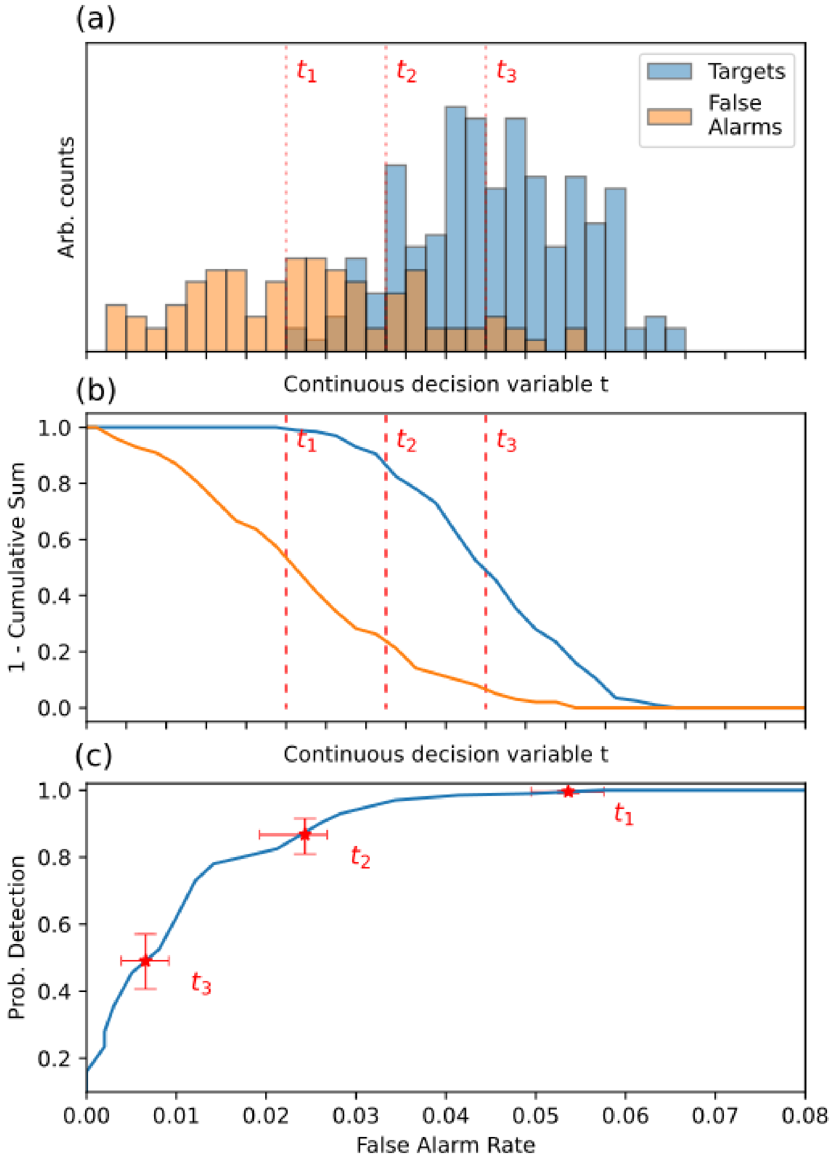


Figure B-14: Receiver Operating Characteristics (ROC) curve is constructed using a varying threshold on a decision variable.

4. By varying the threshold value throughout its range, the resulting PD and FAR can be plotted against one another. This ROC curve is depicted in Figure B-14(c). The mappings of $t_1, t_2,$ and t_3 to the ROC are depicted in order to visualize their relation to the underlying parameter distributions, but this is not typically presented in ROC reported in technical assessments.

5. The ROC curve presents the correlated effect on the PD and FAR due to the variation of an unseen third variable, in this case the decision threshold value. The investigator interprets the ROC to understand the trade-offs between PD

and FAR, which can then lead to a decision on the appropriated decision threshold value to use for the specific requirement. When developing search technologies of their decision algorithms, the universal goal would be to push the ROC curve to the top left; toward 100% PD and 0 FAR. This is not usually achievable, but the ROC can be used to compare sensors or algorithms in an effort to identify those that move most promisingly in this direction.

6. While typically not included in ROC reported in technical assessments, the curve does have an inherit confidence interval, or error, associated with it. Here, two sources of error that may impact the accuracy of the ROC are considered: the experimental resolution of the threshold parameter and the sensitivity of the *1 - Cumulative* distributions to this; and, introduced statistical error if experiments are repeated in order to build the parameter distribution.

7. When a single data set is collected and then analyzed in detail after the fact, such as in Figure B-14(a), then the most significant variability will come from the sensitivity of the decision process to the experimental accuracy of the threshold value. Here, a variation of ± 5 % was applied to the threshold decision value *t*, and the resulting variation on the *1 – Cumulative* distributions are represented by the error bars of Figure B-14(c).

8. There are also times when the appropriate threshold value is fixed during experimentation or measurement and cannot be applied to data after the fact, such as the operating voltage of a sensor element that defines a system’s performance. Instead whole experiments must be repeated with incremental steps applied to this operating parameter in order to build up the data for the ROC.

9. In addition to the sensitivity to the threshold variability, statistical variations must then also be accounted for. If a threshold value increases the PD for a binomial process from $p_1 \rightarrow p_2$, then the resulting data will include the natural statistical variation in observed successes in addition to the additional successes expected from $p_2 > p_1$. The observed FAR would be similarly impacted.

10. Analysis of this convolution is beyond the scope of this introduction, but the scenario is presented in order to highlight the nuances in the construction of ROC curves, and encourage investigators of military search equipment to apply healthy skepticism to reported ROC curves.

11. ROC are most useful in the comparison of very similar technologies on data collected in very similar, if not exact, experimental conditions against very similar, if not exact, target sets. Technical authorities responsible for capability development should be very skeptical of ROC that conflate any variation in sensors, target sets, test locations, or environmental conditions.

12. A very thorough and readable reference on the definition and use of the ROC is available [13], with more advanced treatments addressing the use of ROC in decision support [20].

B.3 REFERENCES

[1] National Institute of Standards and Technology, "NIST/SEMATECH e-Handbook of Statistical Methods," Apr 2012. [Online]. Available: https://www.itl.nist.gov/div898/handbook/index.htm. [Accessed Jan 2020].

[2] M. J. de Smith, Statstical Analysis Handbook, 2018 ed., Chicago, IL: Drumlin Security Ltd, 2018, p. 638.

[3] Wikipedia, "Statistics," Wikipedia, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Statistics. [Accessed Jan 2020].

[4] M. G. Kendall and A. Stuart, The Advanced Theory of Statistics - Distribution Theory, 1958 ed., vol. 1, London: Charles Griffin & Company.

[5] H. a. H. Box, Statistics for experimenters, Wiley, 1978.

[6] R. J. Barlow, Statistics - A guide to the use of statstical methods in the physical sciences, Chichester, UK: John Wiley and Sons, 1989.

[7] L. Brown, T. Caie and A. DasGupta, "Interval estimation for binomial proportion," *Statistical Science,* vol. 16, no. 2, pp. 101-133, 2001.

[8] J. C. Clopper and S. E. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika,* vol. 26, no. 4, pp. 404-413, 1934.

[9] E. Cameron, "On the estimation of confidence intervals for binomial population proportions in astronomy; the simplicity and superiority of the Bayesian approach," *Astronomical Society of Australia,* vol. 28, pp. 128-139, 2011.

[10] E. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association,* vol. 22, pp. 209-212, 1927.

[11] N. Gehrels, "Confidence limits for small numer of events in astrophysical data," *The Astrophysical Journal,* vol. 303, pp. 336-346, 1 Apr 1986.

[12] L. Barker, "A comparision of nine Confidence Intervals for a Poisson parameter when the expected number of events is <= 5," *The American Statistician,* vol. 52, no. 2, pp. 85-89, 2002.

[13] J. Byrne and P. Kabail, "Comparison of Poisson Confidence Intervals," *Communications in Statistics – Theory and Methods,* vol. 34, pp. 545-556, 2005.

[14] V. V. Patil and H. V. Kulkarni, "Comparison of confidence intervals for the Poisson mean: Some new aspects," *REVSTAT – Statistical Journal,* vol. 10, no. 2, pp. 211-227, 2012.

[15] N. C. Schwertman and R. A. Martinez, "Approximate Poisson confidence limits," *Comminications in Statstics – Theory and Methods,* vol. 23, no. 5, pp. 1507-1529, 1994.

[16] J. A. Taylor, An Introduction to Error Analysis, Second ed., Sausalito, CA: University Science Books, 1997.

[17] ASTM International, "Standard test method for estimating limits of detection in trace detectors for explosives and drugs of interest," ASTM International, 100 Barr Harbour Drive, PO Box C700, West Conshohocken, PA USA 19428-2959, 2020.

[18] A. L. Rukhin and D. V. Samarov, "Limit of detection determination for censored samples," *Chemometrics and Intelligent Laboratory Systems,* vol. 105, pp. 188-194, 2011.

[19] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems,* vol. 80, pp. 24-38, 2006.

[20] C.-I. Chang, "Multiparameter Receiver Operating Characteristic analysis for signal detcteion and classification," *IEEE Sensors Journal,* vol. 10, no. 3, pp. 423-442, March 2010.

[21] CEN Workshop Agreement (CWA) 14747, "Humanitarian Mine Action Test and Evaluation - Part I: Metal Detectors," European Committee for Standardization (CEN), 2003.

[22] Government of Canada, "TERMIUM Plus, The Government of Canada’s terminology and linguistic data bank," [Online]. Available: http://www.btb.termiumplus.gc.ca. [Accessed June 2015].

INTENTIONALLY BLANK

ANNEX C - KEY REQUIREMENTS TEMPLATE

ID	User Requirement	Status	Priority	Justification	Threshold MOE	Objective MOE	Remarks
PREPARE							
KUR 1 TRAINING	The User shall be provided with a training provision for the operation and maintenance of the capability.	Candidate	Key	Training should be made available to all Users of the capability.	User provided with the relevant training for their level of interaction with the capability and with training material.	OEM delivers complete training packages including provision of training material.	A [TNA] will be conducted to inform the training solution. Training material could include training aids, training targets, training documentation etc.

INTENTIONALLY BLANK

ANNEX D - BATTLEFIELD MISSION TEMPLATE

1 Introduction

1.1 Background.

1.2 Aim

1.3 Scope

1.4 Objectives

1.5 Limitations

1.6 Success criteria

1.7 Participation

2 Trial scenario and requirements

2.1 Location & Dates

2.2 Equipment

2.3 Key Stakeholders

2.4 Geometry

2.5 Outline scenario description

2.6 Government Furnished Equipment (GFE) list

2.7 Safety zone considerations

3 Execution

3.1 Roles and Responsibilities

3.2 Technical Capability Test Procedures

3.3 Trial site evaluation

3.4 Equipment set up

3.5 System testing and Clutter evaluation

3.6 Ground truth

3.7 Plan of Tests

3.8 Equipment Performance - Live

3.9 Mobility

3.10 Interoperability

3.11 Supportability

3.12 Training

3.13 Main Events List

4 **Service and Support**

4.1 General

4.2 Responsibilities

4.3 Secure Compound

4.4 Accommodation

4.5 Food & Water

4.6 Vehicles

4.7 Fuel

4.8 Testing Site Facilities.

4.9 Power

4.10 Recovery

4.11 Ad Hoc Support

5 **Command and Control**

5.1 General

5.2 Responsibilities

5.3 Trials HQ

5.4 Safety

5.5 Communications

6 **Health and Safety**

6.1 Organizations involved

6.2 Exchange of information

6.3 Health and Safety Plan and Risk Assessments

ANNEX E - BATTLEFIELD MISSION TEST SHEET TEMPLATE

DTG Start		**DTG Finish**	
Team Number			
System No.			

Task Conduct	**Remarks**	**Device Layout**
The conduct of the task is described here and should expand on the diagram in the 'device layout' box.		A visual layout of the target in relation to the ground and any specific features.

SR	**Pri.**	**Threshold MOP**	**Objective MOP**	**R / A / G**	**User Feedback / TM Comments**	**Remarks / Recommendations**
Prepare						
This is a description of the System Requirement which is being trialled.	SR priority.	Description of the Measures of Performance the equipment must achieve to reach the threshold.	Description of the Measures of Performance the equipment must achieve to reach the objective.		Space for user or Trial Manager feedback.	
Project						
Operate						
Protect						
Inform						
Interoperability						
Sustain						

SR	Pri.	Threshold MOP	Objective MOP	R / A / G	User Feedback / TM Comments	Remarks / Recommendations

Faults Details
Users Overall Observations / Comments
Trial Managers Overall Observations / Comments

Data Capturers Signature		**DTG**	

Trial Managers Signature		**DTG**	

ANNEX F - USER QUESTIONNAIRE EXAMPLES

EQUIPMENT X Post Serial Feedback Sheet

To be filled in after each serial

Please Read the Following

The following questionnaire is designed to collect your feedback on serial that you have just completed.

Please answer the questions as thoroughly as you can and provide any additional information that you think is of relevance or would help explain your answers.

Where there are multiple-choice questions please select a single option that most closely reflects your opinion.

If there are questions that do not apply or are to do with functions that you have not used please leave them blank.

Your feedback is much appreciated, thank you.

Please return the completed questionnaire to the trial staff.

Background information

Date: ________________	**Time:** ______________________	
Participant Name / Participant Trial Number		
Serial/Trial Number		
Which Equipment Did You Use?	**MK1**	**MK2**

1. **Please circle one option to show how much you agree or disagree with the following statement: I am confident that the equipment detected all of the targets in the lane?**

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree

Additional Comments:

2. **Please circle one option to show how much you agree or disagree with the following statement: The equipment was easy to use during the serial?**

Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree

Additional Comments:

3. How easy or difficult was it to tell the difference between a target and clutter/background noise?				
1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy
Additional Comments:				

4. How satisfied were you with the speed with which you were able to search?				
1 Very dissatisfied	2 Dissatisfied	3 Neither satisfied or dissatisfied	4 Satisfied	5 Very satisfied
Additional Comments:				

5. Did you have to change the GPR settings during the serial due to a change in ground conditions?	
Yes	**No**

6. How easy or difficult was it to tell that the GPR settings needed to be changed?				
1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy
Additional Comments:				

7. How confident were you that the GPR settings were optimal for the ground conditions once you had changed them?			
1 No confidence at all	2 Not confident	3 Confident	4 Very confident
Additional Comments:			

8. Did you have to change the MD settings during the serial due to a change in ground conditions?	
Yes	**No**

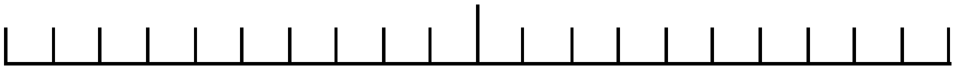
9. How easy or difficult was it to tell that the MD settings needed to be changed?				
1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy
Additional Comments:				

10. How confident were you that the MD settings were optimal for the ground conditions once you had changed them?			
1 No confidence at all	2 Not confident	3 Confident	4 Very confident
Additional Comments:			

Additional Comments and Feedback:

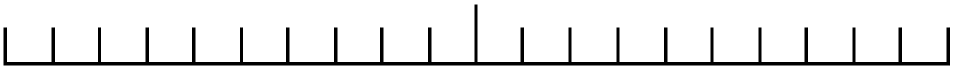
Please place a cross on each of the following lines that reflects your response to each question.

Mental Demand How mentally demanding was the task?



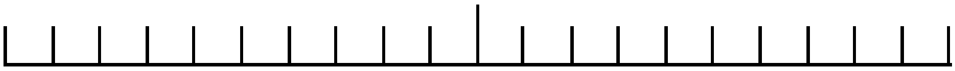
Very Low Very High

Physical Demand How physically demanding was the task?



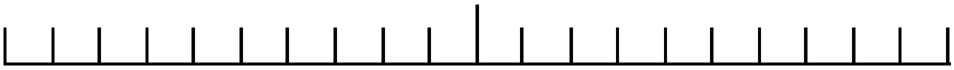
Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?



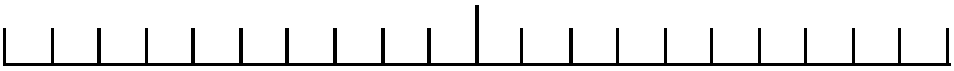
Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?



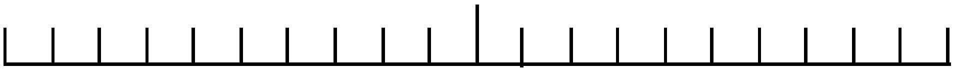
Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?



Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?



Very Low Very High

EQUIPMENT X MK2 End of Trial Feedback Questionnaire

To be filled in at the end of the trial
Please Read the Following
The following questionnaire is designed to collect your feedback on the design and functionality of the EQUIPMENT X MK2 system and how it compares with the in-service (MK1) EQUIPMENT X. Please answer the questions as thoroughly as you can, based on your own experiences during the trial and provide any additional information that you think is of relevance or would help explain your answers. Where there are multiple-choice questions please select a single option that most closely reflects your opinion. If you are filling out a paper version of the questionnaire please ring or put a cross next to your selection, if you are filling this out electronically please delete the other options so that it is clear what your selection is. If there are questions that do not apply or are to do with functions that you have not used please leave them blank. Your feedback is much appreciated, thank you. Please return the completed questionnaire to:

Some information about you	
Participant Name / Participant Trial Number	
What is your current role?	
Roughly how long did you spend using EQUIPMENT X MK2 during the trial?	**______Hrs ________ Mins**

Overall Feedback - Please circle one option to show how much you agree or disagree with the following statements:				
1. I found EQUIPMENT X MK2 easier to use than the EQUIPMENT X (MK1)				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				
2. I think that EQUIPMENT X MK2 was better at detecting targets than EQUIPMENT X (MK1)				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				

3. I think that it was quicker to Search with EQUIPMENT X MK2 than with EQUIPMENT X (MK1)				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				
4. I found it easier to tell the difference between Targets and background clutter (objects that EQUIPMENT X detects, but aren't targets) with EQUIPMENT X MK2 than with EQUIPMENT X (MK1)				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				
5. It was difficult to tell when I needed to change the settings on EQUIPMENT X MK2 to suit a change in ground conditions				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				
6. Overall I would rather search with EQUIPMENT X MK2 than with EQUIPMENT X (MK1)				
Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
Additional Comments:				

Display Characteristics				
7. Overall how easy or difficult was it to view the display while searching (i.e. is the angle and position of the display suitable for use while standing)?				
1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy
Additional Comments:				
8. Overall how easy or difficult was it to read the display in bright sunlight?				
1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy
Additional Comments:				

9. Were there any occasions when you couldn't read the display because of glare or reflections on the display?	
Yes	No
10. Did the display suffer from any scratches since you were using it?	
Yes	No
If yes did this affect your ability to read the information on the display?	
Yes	No

System start-up

11. Was length of time taken for the system to start up (from turning it on to being able to start searching) acceptable?	
Yes	No
Additional Comments:	
12. Was it obvious that the system had started up correctly and was ready to use?	
Yes	No
Additional Comments:	

Pre-sets and adjustment ranges

13. Please provide your opinion regarding the default settings of the following features?

	1 Too Low	2 OK	3 Too High	Did not use / Not Applicable
Display brightness				
MD Volume				
MD Sensitivity				
MD Normal/Mineralised Soil				
GPR Volume				
GPR Surface Removal Setting				
GPR Max Depth Setting				
GPR Shallow Sensitivity				
GPR Deep Sensitivity				
GPR Ground Tracking Algorithm (on/off)				
Additional Comments:				

14. Is the lowest volume setting quiet enough?	
Yes	No
15. Is the highest volume setting loud enough?	
Yes	No
Additional Comments:	
16. Is the lowest brightness setting low enough?	
Yes	No

17. Is the highest brightness setting high enough?

Yes	No

Additional Comments:

MK2 Functionality

18. How useful were the following features?

	1 Not useful at all	2 Slightly useful	3 Useful	4 Very useful	5 Extremely Useful	Did not use
MD Normal/Mineralised Soil						
MD Filter Setting						
MD Sensitivity						
GPR - Surface Removal Setting						
GPR - Max Depth Setting						
GPR - Shallow Sensitivity						
GPR - Deep Sensitivity						
GPR - Ground Tracking Algorithm						

19. How do you think each of the following features affected detection performance?

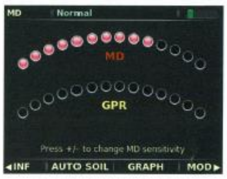
	1 It made it much worse	2 It made it slightly worse	3 It made no difference	4 It made it slightly better	5 It made is much better	Did not use
MD Normal/Mineralised Soil						
MD Filter Setting						
MD Sensitivity						
GPR - Surface Removal Setting						
GPR - Max Depth Setting						
GPR - Shallow Sensitivity						
GPR - Deep Sensitivity						
GPR - Ground Tracking Algorithm						

20. How do you think each of the following features affected the impact of background clutter (objects that EQUIPMENT X detects, but aren't targets) on ground search?

	1 It made it much worse	2 It made it slightly worse	3 It made no difference	4 It made it slightly better	5 It made is much better	Did not use
MD Normal/Mineralised Soil						
MD Filter Setting						
MD Sensitivity						
GPR - Surface Removal Setting						
GPR - Max Depth Setting						
GPR - Shallow Sensitivity						
GPR - Deep Sensitivity						
GPR - Ground Tracking Algorithm						

Using the Display

21. How easy or difficult was it to understand the LED version of the detection strength display?



Press +/- to change MD sensitivity

1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy

Additional Comments:

Using the menus and controls

22. Overall how easy or difficult did you find it to navigate through the menu options on the display?

1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy

Additional Comments:

23. Did you ever get lost in the menu options on the display?

Yes	No

If yes which ones?

24. Overall how easy or difficult was it to understand the information provided on the display?

1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy

Additional Comments:

25. Was it obvious if the system had developed a fault?

Yes	No	Not Applicable

Additional Comments:

26. How easy or difficult was it to change the following settings?

	1	2 Difficult	3	4 Easy	5	Did not use

	Very difficult		Neither difficult or easy		Very easy	
Display brightness						
Volume						
Detection mode (MD Only, MD + GPR etc)						
Detection alert mode (audio only, visual only, etc)						
MD Normal/Mineralised Soil						
MD Filter Setting						
MD Sensitivity						
GPR - Surface Removal Setting						
GPR - Max Depth Setting						
GPR - Shallow Sensitivity						
GPR - Deep Sensitivity						
GPR - Ground Tracking Algorithm - turn on/off						
Additional Comments:						

27. How often did you adjust the following settings?

Settings	After Turning EQUIPMENT X on did you typically change any of the following **(Y/N)**	During a Search typically how often did you change the settings (pick one)?				
		Never	1 to 4 times an hour	5 to 10 times an hour	Every few minutes	Whenever the ground conditions changed
Display brightness						■
Volume						■
Detection mode (MD Only, MD + GPR etc)						
Detection alert mode (audio only, visual only, etc)						■
MD Normal/Mineralised Soil						
MD Filter Setting						
MD Sensitivity						
GPR - Surface Removal Setting						
GPR - Max Depth Setting						

GPR - Shallow Sensitivity						
GPR - Deep Sensitivity						
GPR - Ground Tracking Algorithm - turn on/off						

Additional Comments:

28. Were there any features that you used a lot that were difficult to access?

Yes	No

If yes which ones?

Control (buttons) Design and Layout

29. How easy or difficult was it to reach the controls/buttons from your normal hand position on the handle?

1 Very difficult	2 Difficult	3 Neither difficult or easy	4 Easy	5 Very easy

Additional Comments:

30. Was it obvious what functions each of the controls/buttons performed?

Yes	No

Additional Comments:

31. Were the controls/buttons large enough to operate when wearing gloves?

Yes	No	I Did Not Wear Gloves

Additional Comments:

32. Did the controls/buttons provide enough feedback i.e. could you tell that you had operated them?

Yes	No

Additional Comments:

33. Did you ever accidently press one control/button while trying to use another?

Yes	No

Additional Comments:

Training
34. How long do you think that it would take to train someone that is already a competent user of EQUIPMENT X MK1 to use EQUIPMENT X MK2?

Are there any additional comments or suggestions for improvements that you would like to make?

THANK YOU

INTENTIONALLY BLANK

AEP-4843(A)(1)