Australian Government
Department of Defence
Defence Science and
Technology Organisation

Guidelines for Interpretation of Published Data on Probability of Detection for Nondestructive Testing

C.A. Harding and G.R. Hugo

Maritime Platforms Division
Defence Science and Technology Organisation

DSTO-TR-2622

ABSTRACT

This report provides general guidelines for the interpretation of published data on probability of detection (POD) for nondestructive testing. An overview is provided of the different types of probability of detection data, methods for statistical analysis and the assumptions that may be embedded in these analyses. Four key issues have been identified which need to be addressed when assessing the applicability of published probability of detection trial data to a new nondestructive testing application. Specific consideration should be given to the system boundary, which defines those elements of the inspection process and other factors potentially affecting inspection reliability that are considered to be under examination in the POD trial and those that are considered to be outside the scope of the trial.

RELEASE LIMITATION

Approved for public release

Published by

Maritime Platforms Division
DSTO Defence Science and Technology Organisation
506 Lorimer St
Fishermans Bend, Victoria 3207 Australia

Telephone: (03) 9626 7000
Fax: (03) 9626 7999

AR-015-141
November 2011

Guidelines for Interpretation of Published Literature on Probability of Detection for Nondestructive Testing

Executive Summary

Appropriate application of nondestructive testing (NDT) methods is dependent on knowledge of the minimum sizes of defects that the techniques are capable of reliably detecting, relative to the defect sizes that could be structurally significant. For some applications, the failure of NDT to detect a single defect could cause catastrophic failure including loss of life.

The reliability of NDT is commonly characterised as the probability of detection (POD) of a specific type of defect as a function of defect size. This report provides general guidelines for the interpretation of published data on POD. When probability of detection is estimated using a traditional POD trial in which field NDT technicians perform inspections on specimens with known defects, then the POD information obtained from the trial is strictly applicable only to the exact conditions and defect types for which the POD trial inspections were performed. Any broader application of the estimated POD to other inspection conditions is reliant on an engineering assessment that the change in inspection conditions will not reduce the POD. Four key questions have been identified which are designed to assist engineering staff to assess the applicability of published POD trial data for a new NDT application:

- How closely do the NDT technique and defect and material types used in the POD trial experiment match the new application, and how important are the differences?
- Where were the system boundaries for the POD trial?
- Who conducted the POD trial and for what purpose?
- What has not been said in the reporting of the POD trial results?

The purpose of a POD trial is to obtain an estimate of the POD by acquiring suitable experimental data and conducting an appropriate statistical analysis. Confidence limits are applied to the estimated POD to account for sampling variability inherent in any empirical statistical trial. It is not necessary to have a comprehensive understanding of the methods for statistical analysis of POD data in order to make use of published data. However, an understanding of different types of POD data and assumptions that may be embedded in the analysis methods may be helpful in interpreting the literature.

This page is intentionally blank

Authors

Dr Cayt A Harding
Maritime Platforms Division

Dr Cayt Harding has worked in the nondestructive evaluation (NDE) group of the Defence Science and Technology Organisation since 1999. She completed her undergraduate training at the University of Melbourne, receiving first class honours in Mechanical Engineering and a science degree in Applied Mathematics and Physics. Cayt has twelve years experience in research on reliability assessment of nondestructive evaluation and received a PhD from The University of Melbourne for her research on probability of detection measurement and modelling.

Dr Geoffrey R Hugo
Maritime Platforms Division

Dr Geoff Hugo is the Science Team Leader for NDE research for air vehicle applications. He has a Bachelor of Science and Bachelor of Engineering (with Honours) degrees in Applied Mathematics and Materials Engineering and a PhD in Materials Engineering, all from Monash University. Since joining DSTO in 1986, he has worked on a variety of projects in materials engineering and nondestructive evaluation. Geoff has fifteen years experience in research on advanced computer-based non-destructive evaluation systems, including both ultrasonic and eddy-current systems. He is the author of more than 25 scientific papers and technical reports and is a member of the Institution of Engineers, Australia.

This page is intentionally blank

Contents

ABBREVIATIONS

SYMBOLS

This page is intentionally blank

Abbreviations

ADF	Australian Defence Force
AFHR	aircraft flying hours
CL	confidence limit
DGTA	Directorate General Technical Airworthiness, ADF
DTA	damage tolerance analysis
LPT	liquid penetrant testing
MAPOD	model-assisted probability of detection
NDE	nondestructive evaluation
NDI	nondestructive inspection
NDT	nondestructive testing
NDTSL	Nondestructive Testing Standards Laboratory, DGTA
POD	probability of detection
pdf	portable document format
RAAF	Royal Australian Air Force
SBI	safety by inspection
USAF	United States Air Force

Symbols

a	defect size
a_{crit}	critical defect size
a_{NDI}	minimum reliably detectable defect size
a_{90}	defect size having 90% probability of detection
$a_{90/95}$	defect size having 90% probability of detection demonstrated with 95% statistical confidence
r	NDT response
ε	noise term
β_0	intercept parameter in model of quantitative NDT response as a function of defect size
β_1	gradient parameter in model of quantitative NDT response as a function of defect size
δ	standard deviation of noise term in model of quantitative NDT response as a function of defect size
Ψ	probability of detection
$\hat{\Psi}$	estimated probability of detection
Ψ_L	lower confidence limit on probability of detection

This page is intentionally blank

1. Introduction

Nondestructive testing[1] (NDT) is used to search for defects in structural materials and components, usually for the purpose of assessing whether the material or component is safe or fit for use. NDT is used widely for detection of fatigue cracking and corrosion in metals; porosity, fusion defects and cracks in welds, and disbonds or other anomalies in composite components. NDT methods may also be used to confirm correct assembly of parts or measure component dimensions (e.g. thickness). Some NDT applications are safety-critical, whereas others form part of purely preventative maintenance processes aimed at minimizing more expensive maintenance at a later date.

There are a variety of NDT methods available with differing capabilities. One of the key features that determines appropriate applications of an NDT method is the minimum defect size, a_{NDI}, which can be reliably detected by a technique, relative to the sizes of defects that might be structurally significant. The detectable defect size, and the reliability with which it can be detected, are dependent on many factors, not least of which can be the inherent variability in the characteristics of the defects to be detected. In some cases, for example, inspection of welds, the detectable defect size may be dependant on the specific weld geometry and the specific locations of possible defects within the weld.

Objective knowledge of the reliability of NDT is particularly important for aerospace applications, since NDT (both during production and in service) is a key element of structural integrity management and minimum standards for NDT reliability are specified in airworthiness codes[2]. Failure of NDT to detect a defect may have a variety of consequences including unavailability of aircraft, increased maintenance costs, or catastrophic failure of safety-critical structure. Studies of NDT reliability are usually focused on avoiding catastrophic failure and demonstrating that the requirements set out in airworthiness standards are achieved.

The reliability of NDT is commonly characterised in terms of the probability of detection (POD, Ψ) of a specified type of defect as a function of defect size, a. As will be discussed in Section 2, quantitative assessment of the reliability of NDT is an essential part of aircraft structural integrity management. Current practices for determining probability of detection require large-scale trials of NDT procedures on representative components to gather data for statistical analysis, which can be prohibitively expensive. To account for sampling variability inherent in any empirical statistical trial, it is normal to apply confidence limits to the estimated POD.

---

[1] Also known as nondestructive inspection (NDI) and nondestructive evaluation (NDE). These terms are regarded as synonymous for the purposes of this report.

[2] The structural integrity management philosophies, standards and requirements for other safety-critical applications, such as in the maritime and nuclear domains, are significantly different to the aerospace domain and will not be considered in this report.

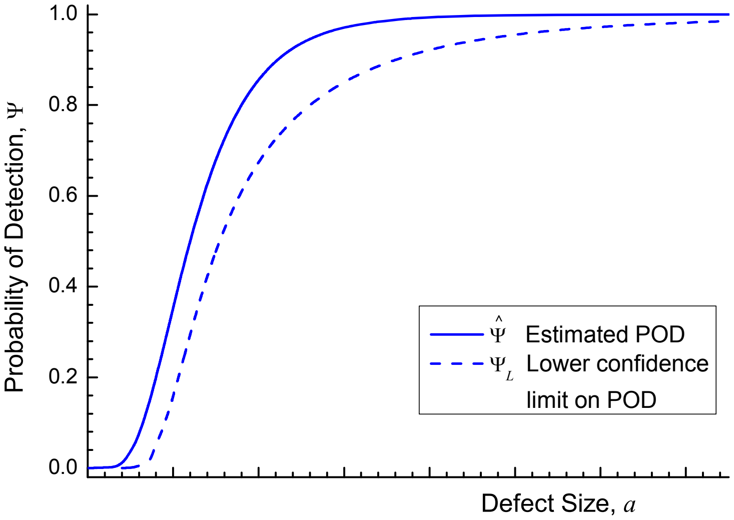

Figure 1 Probability of detection, Ψ, and lower confidence limit on POD, Ψ_L, plotted against defect size, a.

Figure 1 shows a typical estimated POD curve $\hat{\Psi}$ and lower confidence limit Ψ_L, where Ψ_L represents the lower bound on where the true POD curve might lie and still be consistent with the observed data. Two defect sizes are frequently extracted from POD information:

a_{90} is the defect size at which the estimated POD[3], $\hat{\Psi}$, reaches 0.9, i.e. $\hat{\Psi}(a_{90}) = 0.9$, and

$a_{90/95}$ is the defect size at which the lower 95% confidence limit Ψ_L reaches 0.9, i.e. $\Psi_L(a_{90/95}) = 0.9$.

This report provides general guidelines for the interpretation of published literature on POD, as applicable to NDT of ADF aircraft. The purpose is to provide engineering staff, including those within the RAAF Nondestructive Testing Standards Laboratory (NDTSL), with information to assist with the evaluation of limitations for standard NDT methods. In this context, “limitations” refers to the sizes and types of defects that will be reliably detected by an NDT procedure [1].

2. Probability of Detection Requirements for Aircraft Structural Integrity

The damage-tolerance philosophy for aircraft design and certification, also known as safety-by-inspection (SBI), is based on a damage tolerance analysis (DTA), which assesses the ability of the structure to withstand service loads and usage in the presence of damage. Damage tolerance assumes that damage may exist undetected in the structure following production or in-service inspection. The DTA will evaluate the

---

[3] The caret (^) denotes a statistically estimated quantity.

growth rate of a defect (typically a fatigue crack) as a function of aircraft flying hours (AFHR) and also determine the critical defect size for a particular location, Figure 2. A 'safe' inspection interval for SBI management is determined as a prescribed fraction (typically half) of the time in AFHR it takes for the assumed defect to grow from the minimum detectable defect size, a_{NDI}, to the critical defect size, a_{crit}, at which the structure could fail under service loads. A DTA for any given location requires extensive engineering analysis to determine the crack growth rate and critical defect size, often involving the development of detailed finite element analyses and load models applicable to the local area.

The definition of a_{NDI} as the "minimum detectable defect size" can cause confusion and miscommunication between structural integrity engineers and NDT personnel. From the engineer's perspective, a_{NDI}, is the smallest defect size used in their analysis because it is the defect size assumed to be already present in the structure. The analysis predicts the defect growth from that size. Thus for the DTA engineer, a_{NDI} is a minimum defect size that needs to be considered in the analysis. From the NDT perspective, a_{NDI} needs to be the largest defect size that could conceivably remain undetected in the structure following an inspection. a_{NDI} must therefore be the largest defect that might possibly be missed by the inspection under adverse conditions. However, the term "minimum detectable defect size" could mistakenly be interpreted as the smallest defect that could possibly be detected by the method under ideal conditions, which could result in greatly underestimating a_{NDI} and thus compromising the DTA certification. The best textual definition of a_{NDI} is the "minimum *reliably-detectable* defect size", where the definition of "reliably detectable" is elaborated below.

2.1 NDT Reliability Specifications in Aircraft Structural Integrity Standards

For aircraft with an airworthiness certification based on safety-by-inspection, airworthiness standards specify the defect size that is appropriate for use as a_{NDI}. JSSG-2006 *Joint Service Specification Guide Aircraft Structures* is the multi-service guide to the specification of Aircraft Structures for use within the USA Department of Defence [2]. JSSG-2006 includes specific defect[4] sizes that shall be assumed to exist initially in the structure as a result of the manufacturing process, normal usage and maintenance, and following an in-service inspection. For in-service inspections, JSSG-2006 specifies:

> *"The smallest damage which is presumed to exist in the structure after completion of a depot or base level inspection should be as follows unless specific NDI procedures have been developed and the detection capability quantified."* (JSSG-2006, paragraph A.3.12.1 f.) [2]

This paragraph goes on to list flaw sizes that may be assumed for several NDT techniques in a given type of structure. For example: *"The minimum assumed flaw size at locations other than holes should be a through-the-thickness crack of length 0.50 inch when the material thickness is equal to or less than 0.25 inch For material thickness greater than 0.25 inch, the assumed initial flaw should be a semicircular surface flaw with length equal to 0.5 inch and*

---

[4] Defects are also referred to as flaws or damage.

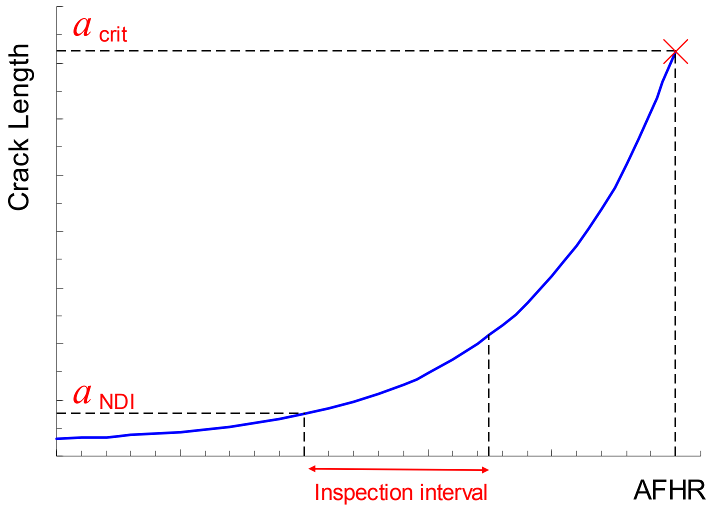

Figure 2 Inspection interval determined from a_{NDI} and crack growth curve (schematic)

depth equal to 0.25 inch." However, JSSG-2006 endorses the use of initial flaw sizes smaller than the standard values subject to a demonstration of the reliability of the NDT process. Specifically, paragraph 4.12.1 states:

> *"Where initial flaw assumptions for safety of flight structures are less than those of 3.12.1, a non-destructive inspection demonstration shall be performed. This demonstration shall verify that all flaws equal to or greater than the assumed flaw size will be detected with a statistical confidence of ___________ ."* (JSSG-2006, paragraph 4.12.1.a) [2]

The blank is intended to be completed by the specification writer for a particular aircraft based on the verification notes for this paragraph. The recommended level of reliability to be demonstrated is given in JSSG-2006 Appendix A which states:

> *"A flaw size smaller than the design flaw size must have a probability of detection of 90 percent. This capability must be verified with a 95 percent confidence level by conducting a statistically valid demonstration."* (JSSG-2006, paragraph A.4.12.1.a Verification Guidance) [2]

Thus under JSSG-2006, the recommended value for a_{NDI} is considered to be the defect size for which a 90% probability of detection has been demonstrated with 95% statistical confidence, commonly denoted $a_{90/95}$. This is the default standard for all damage tolerance analyses of airframe structure for US-built military aircraft.

The US Department of Defense Handbook for Engine Structural Integrity Program, MIL-HDBK-1783B, contains guidance for detectable defect sizes to be used in the management of engine components [3]. MIL-HDBK-1783B generally requires 90% probability of detection to be demonstrated with 95% statistical confidence. However, for some automated inspection systems, MIL-HDBK-1783B allows the best estimate of the defect size having 90% POD to be used instead of the 95% confidence limit value. This a_{90} estimate is described in MIL-HDBK-1783B as the defect size having 90% POD demonstrated with 50% confidence ($a_{90/50}$) and its use is allowed on the basis that an automated inspection is not subject to technician-to-technician variability:

> *"The 90% POD/50% CL requirement can be used for some automated NDI methods based on the NDI process being in control. … Operator variability is the most influential single variable on reliability demonstrations/testing. With the introduction of enhanced automated eddy current inspection systems, the POD/CL requirement was changed to 90% POD/50% CL to reflect the reduced/removed operator variability. However, demonstration of flaw size detection reliability should be required to ensure the system is a controlled process."* (MIL-HDBK-1783B, paragraph A.4.8.2 Requirement Guidance) [3]

Although MIL-HDBK-1783B allows the use of $a_{90/50}$ (the best estimate of a_{90}) rather than the 95% confidence limit value $a_{90/95}$ in engine structural integrity management, the argument provided is not sound. The purpose of the confidence limit applied to the POD estimate is to allow for the (unknown) sampling error inherent in estimating the POD from a finite sample of experimental data, and not to account for variability in the NDT process (e.g. due to human factors). The reduced variability in an automated inspection system will likely result in a steeper estimated POD curve (due to less scatter in the NDT measurements relative to the accept/reject threshold), but confidence limits are still required to account for sampling variability in the estimate. MIL-HDBK-1783B provides a table of minimum initial flaw sizes which are explicitly stated to have 90% POD with 95% confidence for all manual NDT methods.

RAAF practice requires the airworthiness of a particular type of aircraft to be certified against an accepted standard, referred to as the certification basis for that aircraft. Most frequently, the certification basis is the airworthiness standard to which the aircraft was originally designed and manufactured. For example, the RAAF F-111 aircraft was certified for fatigue against US MIL standard MIL-A-83444 (1974) "Airplane Damage Tolerance Requirements" [4]. MIL-A-83444 is effectively a predecessor to JSSG-2006 and was the first USA military publication to specify requirements on demonstration of NDT reliability:

> *"Smaller initial flaw sizes than those specified above may be assumed subsequent to a demonstration, described in 4.2, that all flaws larger than these assumed sizes have at least a 90 percent probability of detection with a 95 percent confidence level."* (MIL-A-83444 paragraph 3.1.1.1.a) [4]

The applicable UK Ministry of Defence Standard, DEF STAN 00-970 "Design and Airworthiness Requirements for Service Aircraft" specifies a different approach to JSSG-2006, in that under DEF STAN 00-970 aircraft are normally certified and managed on the basis of safe-life rather than damage tolerance [5]. However, inspection-based substantiation of serviceability is used for components that are susceptible to defects or damage in manufacture or service. It may also be used to extend the life of selected safe-life components. It is important to note that DEF STAN 00-970 sets inspection intervals by dividing the inspectable life by a factor of 3 (c.f. the factor of 2 typically used or implied in US standards).

The original Issue 1 of DEF STAN 00-970 (1987) mandated a minimum overall probability of detecting a defect before it propagates to critical size in the anticipated usage [6]. However, the current issue (Issue 2) is much more general in the minimum requirement for NDT reliability:

> *"As a general rule, the aim should be to choose a detectable crack size that is very unlikely to be missed at the given location under service conditions. This choice must be guided by experienced NDI operators using accumulated evidence for the technique in question and taking account of the standards that have been achieved when special trials have been done."* (DEF STAN 00-970 Issue 2, Part 1 Section 3 Leaflet 36, paragraph 3.2) [5]

The less stringent detectability criteria found in the more recent DEF STAN 00-970 Issue 2 may reflect the reality that, for many NDT procedures, no reliability demonstration is actually carried out.

2.2 Currently Accepted NDT Procedure Limitations

Resource constraints, combined with the very significant time and effort required to prepare test specimens and conduct POD trials, dictate that, notwithstanding the airworthiness specifications outlined above, experimental POD trials are generally not conducted for individual NDT procedures. The more common approach to determining a_{NDI} for use in a damage tolerance analysis is to rely on an estimated 'limitation' for the technique, which is the smallest defect that a published NDT procedure is expected to reliably find.

> *"NDT Procedure limitations state the type and size of the defect the procedure will readily detect. Limitations are intended only as a guide to engineering staff to assist in the determination of test intervals or the safe working life of an item."* (AAP 7001.068(AM1) paragraph 20) [1]

The limitation is determined based either on laboratory experiments applying the technique to simulated defects (such as machined notches) or, more frequently, from previously accepted values for similar inspection procedures and previous experience with the NDT technique. Limitations are generally not derived using statistical analysis of experimental data. In the RAAF, NDT procedures are developed for specific applications by qualified NDT technicians with extensive practical NDT experience but limited (or no) formal training on reliability issues. The difficulty in adequately addressing probability of detection is acknowledged in the ADF Design and Technology Services Support Manual chapter for the RAAF Non Destructive Testing Standards Laboratory (NDTSL) [1].

Default limitations for each of the standard NDT methods commonly used on ADF aircraft are specified in the general procedure for each method [7]. DSTO is undertaking a series of literature reviews to specifically address POD for a number of the standard NDT methods. A review of the literature on POD for liquid penetrant testing was the first of these to have been completed [8].

3. Probability of Detection Trials

The previous section discussed the need for information on the reliability of NDT procedures used on ADF aircraft. When probability of detection is estimated using a traditional POD trial in which field NDT technicians perform inspections on specimens with known defects, then the POD information obtained from the trial is strictly applicable only to the exact conditions under which the POD trial inspections were performed. Any broader application of the estimated POD to other inspection conditions is reliant on an engineering assessment that the change in inspection conditions will not reduce the POD. This section considers what information is required in order to perform that engineering assessment of whether and how the results of a POD trial described in the literature may be translated to either the general application of that method, or to a specific inspection procedure.

Usually a POD trial is an experiment where the defect size is an independent (controlled) variable and the inspection result (hit/miss or response, *r*) is the dependent variable. The effects of factors, other than defect size, that influence the POD can be incorporated in (or excluded from) a POD trial by:

(i) fixing the factor to a single value or specification that is representative of the field inspections, e.g. limit equipment used to be a specific type, which then limits the applicability of the POD results, or

(ii) randomising the factor from within a pool of possible conditions that are representative of the field inspections, e.g. conduct the POD trial using a range of inspectors drawn from the population who normally conduct the inspections, or

(iii) explicitly controlling factors using a formal design of experiments, so that the effect of these factors can be quantitatively examined. This might be most appropriate for easily controlled or well defined factors such as probe size or frequency.

POD trials vary enormously in both scope and purpose. These range from large-scale trials intended to benchmark the reliability of field NDT performed across the entire USAF, to laboratory trials intended to compare the intrinsic capabilities of different equipment or technologies for a particular inspection scenario. Two key elements, which are defined by scope and purpose of a trial, are the nature of the specimens and the boundary of the system to be considered.

For some types of inspections, such as inspection of turbine engine disks, adequate numbers of ex-service components containing real in-service defects may be available for use in POD trials. However, this is the exception rather than the rule and for airframe inspections it is extremely rare to have real components available which contain in-service defects of a size suitable for use in POD assessment. Instead, specimens are generally manufactured specifically for a POD trial and simulated defects are introduced into a proportion of the specimens. The fidelity of both the specimens and the simulated defects to represent “real” in-service conditions and defect characteristics varies enormously for different POD trials. The cost of specimen fabrication and defect insertion escalates exponentially with fidelity.

3.1 System Boundaries in POD Trials

For a POD trial, the system boundary defines which elements of the inspection process, and which of the other factors potentially affecting inspection reliability, are considered to be under examination in the POD trial and which are considered to be outside the scope of the trial. The definition of the system boundary is frequently the most difficult information to infer when reviewing published results for POD trials conducted at other laboratories. As an example, if technicians failed to find defects because they inspected the wrong specimen, used the wrong procedure, or reported the results incorrectly, would that have been treated as a miss under the protocol for the trial? Such negative results might be excluded from the data during analysis on the basis that they were caused by factors outside the scope of the NDT process, as defined for the conduct of the POD trial. In some cases, the experimenter does not consider the system boundary explicitly, and it is only implicitly defined by the purpose of the trial and by the environment within which it was conducted.

For some POD trials, the system being assessed in the experiment is limited to the interrogation signal (e.g. an ultrasound beam) directly interacting with a defect to give a response. It is assumed that the equipment is calibrated and used correctly, and that the interrogating signal actually encounters the defect. In this case, only the intrinsic capability of the inspection method is being measured by the POD trial and human factor issues in operating the equipment or geometry issues, such as whether the probe actually passes over the defect, are excluded from the trial. This type of POD exercise may be very useful for improving aspects of the inspection process or comparing different settings or different equipment, but is likely to be of limited or no value for assessing the overall POD for the NDT procedure as applied in the field.

At the other end of the spectrum, some POD trials are intended to determine the probability of detecting defects in a particular component using a fielded inspection procedure, taking into account all possible real-world causes of a defect being missed. As with specimens, the more accurately the trial conditions reflect the reality of field inspection conditions, the greater the cost.

Typically, the more comprehensive a POD is designed to be, in terms of capturing as many elements as possible within the system boundary, the more application-specific it becomes and the more difficult it becomes to translate the final POD results across to other applications. By comparison, POD results from a trial that *excludes* all factors other than the intrinsic variation in the equipment are probably applicable to most inspections that use that type of equipment. However, the difficulty with making use of those results to predict field NDT performance is that other causes of failure to detect defects (beyond the intrinsic capabilities of the equipment) — such as poor test area coverage or variations in defect characteristics — will not have been considered.

If measurement of false call rates is to be attempted, then defining the system boundary is particularly important. For field inspections, there are usually a myriad of possible engineering or maintenance actions when a defect is detected, including repeating the inspection on the spot, applying a more sensitive backup procedure, or polishing or reworking the area before repeating the inspection. A positive NDT indication may trigger a hierarchy of actions which have increasing cost for the maintainers. This hierarchy allows for some incidence of false calls at each different level in the process;

the significance of the false call rate increases at each level as the cost to resolve the problem increases. Usually it would only be feasible to address the lowest levels of corrective action (e.g. repeat the inspection, or perhaps apply a backup procedure) within the scope of a POD trial.

4. Sources of Literature

4.1 Conference Presentations and Conference Proceedings

Conference presentations (PowerPoint slides) are generally the least comprehensive source of information about POD trials. Results presented in slides are often selectively chosen to best illustrate the author's main points and, due to the limited space on the slides, caveats or limitations that relate to the information presented may be omitted. For written papers that are formally published in conference proceedings, there is more scope for the author to fully explain the data presented, but it is to be anticipated that the results may still be selectively chosen to illustrate the key points that the author wishes to make. Conference papers also often report on research that is still in progress or even only in the early stages. Thus, conference papers might provide only a partial picture of the results obtained to date and not the final results of the completed study, or they may give only an incremental update on results presented previously. Conference papers are often not peer-reviewed, meaning there is no independent evaluation of the information presented in the paper, either in terms of the clarity of presentation or the validity of the conclusions relative to the data presented.

Not withstanding these caveats, some useful conference papers on NDT reliability and POD studies can be found at:

- http://www.ndt.net/
- http://www.jcaa.us/
- http://www.cnde.iastate.edu/QNDE/pastconferences.htm

4.2 Journal Articles

Journal articles are usually subjected to a peer review prior to publication. This means that other experts in the field are requested by the journal editors to comment on the paper, including aspects such as the originality and significance of the research, the clarity of presentation, and the validity of the conclusions based on the data presented. The peer review process minimises the dissemination of irrelevant findings or unwarranted claims and helps maintain the integrity of the journal.

Journal articles are usually written once the research is completed and clear conclusions can be drawn from the results. Consequently, they should give a more complete picture of a project that might be expected in a conference paper. However, journal articles may be restricted in length, which may preclude the inclusion of substantial detail about the conduct of POD trials. They are often written for a broad audience and may only include illustrative results, which would of course be the best examples to support the authors' conclusions.

4.3 Reports

Formal reports published by the organisation that undertook the research are usually the best source of information on a specific POD trial. They usually contain a comprehensive description of the experiment, including descriptions of specimens, qualification levels of participants and design of experiments. A good report should enable a subject matter expert to make an informed assessment of the overall quality of the trial and the associated data analysis. It will clearly define the boundaries of the system under consideration and provide enough information for a reader to make an assessment of how the results translate to other applications.

Some formal reports are written for a very specific audience and assume a high level of background information about the project, in which case other related publications such as conference papers may be helpful for understanding the context in which the POD trial was conducted. However, the greater detail in formal reports can be extremely useful in assessing how the results may be translated to other applications.

Formal reports can be more difficult to obtain than other forms of published literature and there may be a considerable gap between the completion of the research and publication of the final report. Citations of formal reports in other documents will include a report number, which makes it considerably easier to locate the report. Some online repositories make reports available in pdf format for free download, particularly reports from government-funded research projects. See for example:

- http://www.dsto.defence.gov.au/publications/
- http://www.dtic.mil/dtic/

4.4 Standards and Handbooks

Some NDT probability of detection information can be found in standards and handbooks. Generally, the detectable defect sizes given in standards are intended to be conservative sizes, being the largest defects that might be missed under 'normal' operating conditions. However, many of the "standard values" quoted for detectable defect sizes are "historically accepted" values, which may or may not be underpinned by reliable data. For example, an extensive DSTO review of documentary evidence related to POD for magnetic rubber inspections failed to find sufficient documented empirical justification for the value of a_{NDI} = 0.020 inch commonly accepted as reasonable for this method [9]. (A POD trial was subsequently undertaken by DSTO which supported the validity of the 0.020 inch value for active-field magnetic rubber inspections [10].)

5. Key Questions for Evaluation of Published Data

There are four key questions that should be kept in mind when evaluating published POD data:

- How closely do the NDT technique and defect and material types used in the POD trial experiment match the new application, and how important are the differences?
- Where were the system boundaries?
- Who conducted the POD trial and did they have a specific agenda?
- What has *not* been said?

These questions are elucidated in the sections below.

5.1 How closely do the NDT technique and defect and material types used in the POD trial experiment match the new application, and how important are the differences?

This question addresses the technical similarity of the NDT methods used in the POD trial to the intended application. This requires critical analysis to assess which aspects of the inspection would have the biggest impact on the reliability and are therefore the most important. Aspects to consider include:

- the nature of the defects to be detected,
- the material, surface finish and local geometry of the part, and
- the inspection technique, including variations in equipment, inspection parameters (e.g. frequency), and calibration processes.

It is important to beware of the academic researcher who refers to "cracks" when their experiments actually used artificially machined notches, or some other manufactured discontinuity, to simulate cracking.

It is often easier to identify the differences than it is to evaluate their importance. At best, it might be possible to assess whether a particular difference between the conditions used for the POD trial compared to the new application would lead to an over- or under-estimate of POD. There may be multiple differences, some of which would tend to over-estimate the POD and some which would under-estimate it, and it would usually be very difficult to weigh up these effects in the absence of any supporting quantitative evidence.

Recent progress in NDT reliability research has been in the area of model-assisted probability of detection (MAPOD) assessments [11 , 12 , 13]. MAPOD uses models of the underlying inspection process to assist with predicting the probability of detection for an inspection, possibly incorporating data from a variety of sources and employing physics-based modelling of the inspection process where possible. One benefit of model-assisted approaches to POD assessment is that models can provide tools to quantitatively consider the effect of specific factors on the overall POD and this has the potential to increase the portability of POD information across related applications.

5.2 Where were the system boundaries?

The importance of system boundaries was discussed in Section 3.1. The types of questions that help identify the system boundaries used for a POD trial are:

- Were the inspections performed by NDT technicians from the relevant field or production line environment, or were they performed by specialist laboratory staff or researchers? Were the inspections performed blind, or could there have been some prior knowledge of the type, number or locations of defects? These questions help establish the extent to which extent human factors are incorporated in the trial results.

- What constitutes a miss and what constitutes a hit? For example, if the wrong part was inspected or an incorrect inspection procedure was used, is that considered invalid data (and therefore excluded from the POD data set), or was it included as a realistic possibility for field inspections?

- How much of the field inspection process was captured in the trial? Was the reporting process representative?

It is very difficult to establish defect reporting processes for use during a POD trial that are typical of field inspections, because technicians will usually encounter many more defects during a POD trial exercise than they would in routine field inspections. Consequently, reporting processes which are fully representative of field NDT practice can become very onerous: the reporting may then distract from the technicians' primary role of inspecting, or the technicians may make improvised shortcuts in the reporting process potentially degrading the quality of the POD data.

5.3 Who conducted the POD trial and for what purpose?

It is important to consider the background of the organisations or individuals who designed and completed the POD study. What was their motivation for conducting the trial? This may be a source of potential bias in the results.

Some equipment vendors conduct POD trials to demonstrate the performance of their equipment, which can provide very useful data to underpin future application of that equipment. However, these trials may be structured to demonstrate the strengths of the equipment and it is the job of the consumer to look also for the weaknesses. Because of the high cost of specimen fabrication, such POD trials sometimes utilise the very same specimens that were used previously during the development of the equipment and/or the associated inspection procedures. This is potentially a very serious source of bias towards overestimating the field POD, as the system will have been optimised to find those particular defects. The bias may be particularly severe if the defects in the specimen set used for the POD trial encompassed only a narrow subset of the full range of defects likely to be encountered in the field.

A frequent strength of POD trials performed by equipment vendors is that they have a good understanding of the field environment for the inspection. By contrast, some academic researchers demonstrate relatively little understanding of the practical difficulties in conducting a field inspection, particularly the importance of considering representative complex geometries and surface conditions. It is difficult to translate POD results obtained from simulated defects at the centre of a small flat plate to field inspections of large, complex components.

5.4 What has *not* been said?

Valuable insight can often be obtained by identifying what information has not been provided. For example, if a report makes no mention of the background of the technicians involved in a POD trial, then it is quite possible that inspections were actually performed by "expert users" (for example, technical specialists employed by an NDT equipment manufacturer) who may not be representative of typical field NDT technicians. Making POD trials representative of real inspections is usually difficult and expensive, and so publications reporting on POD trials that have taken these issues seriously usually discuss how the challenges of representing real-world inspections were actually addressed. If these issues are not discussed in a published article, then conservative assumptions should be made, i.e. assume that the estimated POD is higher than would be achieved by field inspections.

5.5 Other Relevant Questions

Consideration of the above questions provides a good starting point for understanding the strengths and limitations of a POD trial based on published information. However, they are far from exhaustive. Some other valuable questions and important issues to be considered include:

- How did the researcher establish the true size of the defects? It is usually difficult and/or expensive to conclusively determine the size of the defects used in a POD trial. If defect size was estimated using nondestructive methods then, as a minimum, fractographic examination of a sample of defects should have been used to establish the accuracy of the sizing method and reveal any systematic bias in the estimated sizes.
- A serious weakness in some POD trials occurs when the "true" set of defects contained in the specimen set is simply assumed to comprise all of the defects found by any of the participating technicians during the POD trial. This assumption has sometimes been made when a POD trial has been conducted on ex-service components which were not destructively examined after the trial to determine the complete defect population. It has the potential to greatly over-estimate the true POD by excluding from the analysis defects that were missed by all technicians. It could also lead to an under-estimation of POD if a false call made by one technician is treated as a defect missed by the other technicians.
- Was the reporting threshold set lower for the POD trial than it would be in practice for field inspections? This is not at all uncommon. An equipment vendor or participating technician may 'turn up the gain' on their equipment, or reduce the reporting threshold, in order to minimise the number of defects they miss, knowing that any false calls on the POD specimens will not incur a cost or maintenance penalty. A high false call rate is a good indicator that the threshold

used during the trial was not representative of what would be realistic for field inspections.

6. Approaches for Statistical Analysis of POD Data

It is not necessary to have a comprehensive understanding of the methods for statistical analysis of POD data in order to make use of published data. However, there are a few key concepts that may be helpful in interpreting the literature.

6.1 Hit/Miss POD Data and *r* vs *a* POD Data

A conventional POD trial involves a large number of inspections conducted on a set of specimens containing known defects.[5] For each inspection of each defect, either a hit or a miss is determined from the inspectors' inspection results (hit/miss data) or, in some cases, a quantitative response (r) is recorded which can be correlated to defect size (r vs a data).[6] This quantitative response is usually an output from the inspection equipment, such as a voltage, signal amplitude, observed defect length, or area. Quantitative response POD data provides more information about the inspection process than hit/miss data. However, determining POD from r vs a data makes assumptions about the mechanisms by which a defect might be missed and requires an explicit definition of the response (or reporting) threshold above which a defect will be detected. An r vs a analysis usually assumes that if the response exceeds the defined reporting threshold it will always be detected. This does not allow for the possibility of the response not being observed by the technician despite the fact that it exceeds the threshold. The relevance of these possibilities will depend on the type and configuration of NDT equipment being used.

Hit/miss data are by far the most commonly available type of POD data. For some NDT techniques, such as radiography, penetrant or magnetic particle inspections, it may be difficult to specify a simple quantitative measurement which primarily determines detection based on comparison to a set threshold. Even for techniques such as ultrasonics or eddy current which readily give a measurable scalar response from the defect, hit/miss data may still give a better representation of the overall performance of a field inspection. This is because hit/miss data may capture human factors involved in set up, calibration and operator interpretation of the data which might be excluded from analysis of r vs a data. The researcher's choice to use r vs a or hit-miss data may influence the explicit or implicit definition of the system boundaries for the POD trial, as discussed in Section 3.1 above.

---

[5] Defect locations and sizes are known to the POD trial organisers, but not to the participating technicians.

[6] This is commonly referred to in the literature as $\hat{a}$ vs a data, where $\hat{a}$ is a signal strength that may be correlated with defect size, a. In this report, r is used for NDT response rather than $\hat{a}$, as the accent ^ is reserved for estimated quantities.

Analysis of POD data needs to consider the types of defects which need to be detected and whether they are expected to have the same or different POD curves. If different, the POD data needs to be grouped for analysis into the appropriate categories of defects and inspection conditions, with a separate POD curve determined for each. An analysis which inappropriately pools the data for different defect types (or different inspection conditions) will generally result in an 'averaged' POD curve which would be flatter (rise less sharply with defect size) than any of the individual POD curves and thus may underestimate the overall POD for large defects. This resultant averaged POD may not be truly representative of any of the different categories of defects.

A contrasting problem occurs if the POD data are sub-divided inappropriately into too many categories of defects or inspection conditions. This will result in a large number of separate POD curves, each based on a relatively small data set giving greater scatter in the POD estimates. This can also create difficulties for engineering interpretation to assess an overall a_{NDI} value.

6.2 Estimating POD

Every NDT system has an actual true probability of detection of defects of a given size and type, whose exact value is unknown. The purpose of a POD trial is to obtain an *estimate* of the POD by acquiring suitable experimental data and conducting an appropriate statistical analysis. Most methods for estimating POD make some assumptions about the form of relationship between POD and defect size. It is also possible to estimate POD as a function of some variable other than defect size, but knowing POD as a function of defect size is usually the most important information because it relates the performance of the NDT system to the structural integrity of the component.[7]

Analysis methods for hit/miss POD data fall into two main categories, interval methods and curve fitting methods. Interval methods group the available data into defect size intervals, and then apply binomial sampling statistics to determine a POD that applies to each size interval. This provides a "step-wise" estimate of POD as a function of defect size. Curve fitting methods assume a suitable mathematical function to describe the POD relationship with defect size and then adjust the free parameters in the chosen function to find the best fit of the chosen functional form to the experimental POD data. POD curve fitting applied to hit/miss POD data is the most common method used for modern POD data analysis and is recommended in guidance publications such as MIL-HDBK-1823A *Nondestructive Evaluation System Reliability Assessment* [14]. Older methods such as the binomial interval methods (and the related 'optimised probability method') are considered to be obsolete and no longer best practice for most POD data analyses.

Analysis of r vs a POD data requires a model of the quantitative NDT response, r, as a function of defect size, a, and incorporates a noise term, ε, which has a random probability distribution. For example,

---

[7] Defect size is usually the most important parameter used by structural integrity engineers to assess the risk that a defect could cause structural failure.

$$\ln r = \beta_0 + \beta_1 \ln a + \varepsilon, \quad \varepsilon \overset{d}{=} \mathrm{N}(0, \delta)$$

where the noise term ε is assumed to have a standard normal distribution with zero mean and standard deviation δ. It is assumed that a defect is detected if the response r exceeds some decision or reporting threshold, in which case it follows that the POD will be described mathematically by a cumulative probability distribution as a function of defect size.

For r vs a analysis, the relationship between NDT response and defect size will generally only hold for a very specific set of conditions. Even within a specific application there may be a number of sub-populations of defects which are governed by different phenomena and therefore have a different response for the same defect size. In addition, the POD curve determined from r vs a analysis may be highly sensitive to the assumed detection threshold, which often cannot be defined independently from the calibration for each inspection. For some procedures, the detection threshold is dynamically adjusted in response to changes in local background noise level. Consequently, there are a number of factors which may complicate an r vs a analysis. By comparison, hit/miss data effectively capture the influence of all these factors, as long as they are represented in the trial.

6.3 Confidence Limits

Using curve fitting methods, it is mathematically possible to estimate POD based on very few data points. However, for such small data sets, the estimated POD could vary significantly from the actual POD due to the sampling variability inherent in any statistical trial. The level of confidence in the accuracy of the POD estimates increases with increasing data set size. It is useful to compute the range or confidence interval within which the true POD might reasonably be expected to lie, given the data set size and the results of the experiment. Confidence intervals always have an associated confidence level, which is usually expressed as a percentage such as a 95% confidence interval. The confidence level defines the likelihood that the computed confidence interval actually contains the unknown true POD.[8] The higher the confidence level, the wider the interval will be, but the greater the confidence that it actually contains the true POD. The larger the sample size (i.e. more inspection data) then the narrower the confidence interval will be for a given confidence level. The lower and upper confidence limits define the lower and upper end points of the confidence interval. For POD curves expressed as a function of defect size, the upper and lower confidence limits define two separate curves lying above and below the estimated POD curve and between which the true POD is expected to lie.

In estimating POD, it is usual to compute a lower confidence limit on the POD which will provide a conservative result when used in subsequent engineering analysis to determine inspection intervals or overall risk of component failure. The upper confidence limit on POD generally has no engineering value and is therefore not

---

[8] For a 95% confidence level, there is a 5% chance of obtaining a data set for which the computed confidence interval *does not* contain the true POD.

computed. Consequently, a one-sided lower confidence limit is normally determined, as shown in Figure 1.

Often for aerospace applications, the statistic of most interest is the defect size at which the POD reaches 90%. A lower confidence limit on POD translates to an upper confidence limit on the defect size for a given POD.

The best estimate of the defect size at which POD reaches 90% is known as a_{90}. The upper 95% confidence limit on the defect size for which the POD reaches 90% is known as $a_{90/95}$. One method to compute $a_{90/95}$ is simply to take the defect size at which the *lower* 95% confidence limit curve reaches 90% POD, as shown graphically in Figure 3. An alternative method to compute $a_{90/95}$ is to directly compute an upper 95% confidence limit on the defect size at which the true POD reaches 90%. This method typically gives a less conservative (i.e. smaller) $a_{90/95}$ value than taking the defect size at which the lower 95% confidence limit curve reaches 90% POD. The lower 95% confidence limit is expected to be conservative with respect to the true POD curve for 95% of all random trials.

Generally, the best-fit POD curve and associated a_{90} defect size provide the best information about the trial results and performance of the NDT method. These statistics are generally robust with respect to the details of the analysis methods used. There is now general consensus amongst NDT reliability practitioners that maximum likelihood estimation is the preferred method for estimating a best-fit POD curve from a POD data set. In contrast, reported confidence limits (lower confidence limit curve and $a_{90/95}$ values) will have been influenced by the size and consistency of the data set, as well as by the analysis method applied. Historically, a number of different methods have been used to compute confidence limits on POD, with different methods potentially giving substantially different confidence limits for the same data set. [9] Consequently, when interpreting POD data from published literature, greater reliance can generally be placed on the reported best estimates of POD (e.g. best fit POD curve and a_{90}) than on confidence limit values (e.g. $a_{90/95}$). Further information on estimating POD from trial data and the application and validity of confidence limits may be found in references [14, 15, 16].

---

[9] Some of the methods used in earlier publications (typically prior to 2001) have subsequently been shown to be invalid for hit/miss POD data, giving unconservative confidence limits on POD [15].

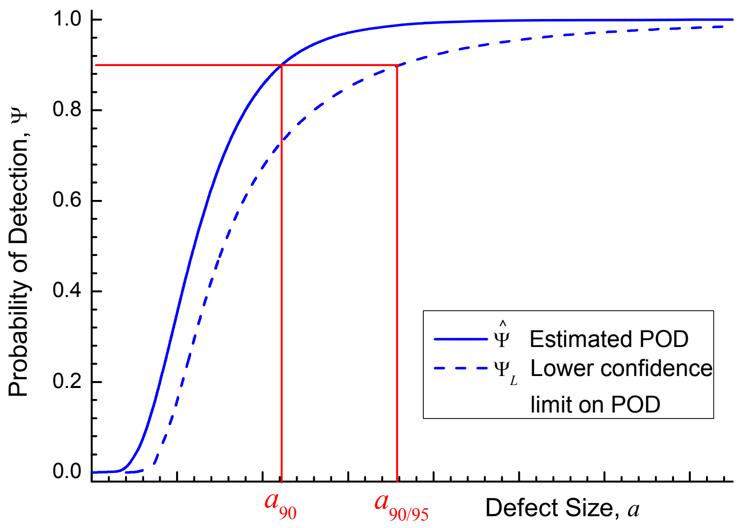

Figure 3 Probability of detection, Ψ, and lower confidence limit on POD, Ψ_L, plotted against defect size, a, showing a_{90} and $a_{90/95}$ values.

7. Conclusions

This report provides general guidelines for the interpretation of published literature on probability of detection for nondestructive testing. When probability of detection is estimated using a traditional POD trial in which field NDT technicians perform inspections on specimens with known defects, then the POD information obtained from the trial is strictly applicable only to the exact conditions under which the POD trial inspections were performed. Any broader application of the estimated POD to other inspection conditions is reliant on engineering judgement that the change in inspection conditions will not reduce the POD.

Four key questions have been identified which are designed to assist a reader to assess the applicability of published POD trial data for a new NDT application.

- *How closely do the NDT technique and defect and material types used in the POD trial experiment match the new application, and how important are the differences?*
  This question addresses the technical similarity of the NDT methods used in the POD trial to the intended application and requires critical analysis to assess which aspects of the inspection would have the biggest impact on the reliability and are therefore the most important.

- *Where were the system boundaries?*
  For a POD trial, the system boundary defines which elements of the inspection process, and which of the factors potentially affecting inspection reliability, are considered to be under examination in the POD trial, and which are considered to be outside the scope of the trial.

- *Who conducted the POD trial and for what purpose?*
  It is important to consider the background of the organisations or individuals who designed and completed the POD study as that may reveal a source of potential bias in the results.
- *What has not been said?*
  Valuable insight can often be obtained by identifying what information has not been provided in published reports.

The purpose of a POD trial is to obtain an estimate of the POD by acquiring suitable experimental data and conducting an appropriate statistical analysis. Confidence limits are applied to the estimated POD to account for the sampling variability inherent in any empirical statistical trial. It is not necessary to have a comprehensive understanding of the methods for statistical analysis of POD data in order to make use of published data. However, an understanding of different types of POD data and any assumptions embedded in analysis methods may be helpful in interpreting the literature.

8. References

1. *Design and Technology Services Support Manual, Section 2 Chapter 9 Non Destructive Testing Standards Laboratory*. (2008) AAP 7001.068(AM1), Australian Defence Force, Australian Air Publication.

2. *Joint Service Specification Guide Aircraft Structures*. (1998) JSSG-2006, Department of Defense, USA.

3. *Engine Structural Integrity Program (ENSIP)*. (2004) MIL-HDBK-1783B, USA.

4. *Military Specification Airplane Damage Tolerance Requirements*. (1974) MIL-A-83444, United States Air Force, USA.

5. *Design and Airworthiness Requirements for Service Aircraft*. (1999) DEF STAN 00-970 Issue 2, Ministry of Defence, UK.

6. *Design and Airworthiness Requirements for Service Aircraft*. (1987) DEFSTAN 00-970 Issue 1, Ministry of Defence, UK.

7. *Non Destructive Testing General Procedures*. (1999) AAP 7002.043-36, Royal Australian Air Force, Australian Air Publication.

8. Harding, C. A. and Hugo, G. R. (2011) *Review of Literature on Probability of Detection for Liquid Penetrant Nondestructive Testing*. DSTO-TR-2623, Defence Science and Technology Organisation, Melbourne, Australia.

9. Hugo, G. R. and Scala, C. M. (2001) *An Assessment of Existing Magnetic Rubber Inspection Probability of Detection Data for F-111 D6ac Steel Structure*. DSTO-TN-0353, DSTO Aeronautical and Maritime Research Laboratory, Melbourne, Australia.

10. Harding, C. A. and Hugo, G. R. (2002) *Experimental Determination of the Probability of Detection for Magnetic Rubber Inspections of F-111 Steel Components*. DSTO-TN-0420, DSTO Aeronautical and Maritime Research Laboratory, Melbourne, Australia.

11. Thompson, R. B. (2008) A Unified Approach to the Model-Assisted Determination of Probability of Detection. *Materials Evaluation* **66** (6) June, pp. 667-673.

12. Knopp, J. S., et al. (2007) Investigation of a Model-Assisted Approach to Probability of Detection Evaluation. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 26, pp. 1775-1782, American Institute of Physics, Melville, New York, USA.

13. Harding, C. A., Hugo, G. R. and Bowles, S. J. (2009) Application of Model-Assisted POD Using a Transfer Function Approach. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 28, pp. 1792-1799, American Institute of Physics, Melville, New York, USA.

14. *Nondestructive Evaluation System Reliability Assessment*. (2009) MIL-HDBK-1823A Department of Defense, USA.

15. Harding, C. A. and Hugo, G. R. (2003) Statistical Analysis of Probability of Detection Hit/Miss Data for Small Data Sets. In: Thompson, D. O. and Chimenti, D. E. (eds.) *Review of Progress in Quantitative Nondestructive Evaluation*. Vol. 22, pp. 1823-1844, American Institute of Physics, Melville, New York, USA.

16. Harding, C. A. (2008) *Methods for Assessment of Probability of Detection for Nondestructive Inspections*. PhD Thesis, The University of Melbourne, Victoria, Australia.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE Guidelines for Interpretation of Published Data on Probability of Detection for Nondestructive Testing	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)
4. AUTHOR(S) C.A. Harding and G.R. Hugo	5. CORPORATE AUTHOR Defence Science and Technology Organisation 506 Lorimer St Fishermans Bend Victoria 3207 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2622	AR-015-141	Technical Report	November 2011

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2009/1136447/1	AIR07/101	DGTA-ADF	20	16

13. DSTO Publications Repository http://dspace.dsto.defence.gov.au/dspace/	14. RELEASE AUTHORITY Chief, Maritime Platforms Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.shtml

Nondestructive testing, Structural integrity, Reliability, Probability.

19. ABSTRACT

This report provides general guidelines for the interpretation of published data on probability of detection (POD) for nondestructive testing. An overview is provided of the different types of probability of detection data, method for statistical analysis and the assumptions that may be embedded in these analyses. Four key issues have been identified which need to be addressed when assessing the applicability of published probability of detection trial data to a new nondestructive testing application. Specific consideration should be given to the system boundary, which defines those elements of the inspection process, and other factors potentially affecting inspection reliability, that are considered to be under examination in the POD trial and those that are considered to be outside the scope of the trial.