Australian Government
Department of Defence
Defence Science and
Technology Organisation

Review of Literature on Probability of Detection for Magnetic Particle Nondestructive Testing

S.K. Burke and R.J. Ditchburn

Maritime Platforms Division
Defence Science and Technology Organisation

DSTO-TR-2794

ABSTRACT

A critical review of the available literature concerning the minimum reliably detectable defect size a_{NDI} for magnetic particle testing (MPT) in aerospace applications is presented. Four probability of detection (POD) studies relevant to detection of fatigue cracks in aircraft components were found over the period 1968 to 2011. As the statistical methods used in these four previous studies were either outdated or otherwise deficient, the original data were reanalysed using currently accepted techniques. A meta-analysis of the results is presented, with emphasis on statistical inferences for the defect size expected to be detected with 90% POD. It is shown that the minimum reliably detectable defect size a_{NDI} = 2.0 mm currently specified by the Royal Australian Air Force for wet fluorescent magnetic particle inspection using the continuous method is consistent with estimates of the average performance of MPT derived from the reanalysis of the literature.

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: 1300 DEFENCE
Fax: (03) 9626 7999

AR-015-495
January 2013

APPROVED FOR PUBLIC RELEASE

Review of Literature on Probability of Detection for Magnetic Particle Nondestructive Testing

Executive Summary

In the damage-tolerance approach to airworthiness, fatigue-critical aircraft structure is subject to regular nondestructive inspection (NDI) to prevent catastrophic structural failure. Periodic inspection intervals are determined using knowledge of the minimum crack size that can be reliably detected, together with information on the structural loads, critical defect size and crack growth rates. The minimum reliably detectable crack size, a_{NDI}, is determined through analysis of the probability of detection (POD) of defects as a function of defect size. Quantitative measurement of a_{NDI} for a given NDI procedure is arguably as challenging to obtain as the other key inputs to damage-tolerance analyses.

The Defence Science and Technology Organisation (DSTO) is conducting a series of critical literature reviews to examine the reliability of standard non-destructive inspection methods used for Australian Defence Force (ADF) aircraft. The first review (DSTO-TR-2623) examined the reliability of liquid penetrant testing (LPT). The present report is the second in the series and is concerned with magnetic particle testing (MPT). MPT is a technically mature inspection method used to detect surface-breaking cracks in high-strength steel components.

A survey of the available literature on the reliability of MPT found some twenty references relevant to POD over the period 1968 to 2011. After critical review, four published studies were considered applicable to detection of fatigue cracks in high-strength steel aerospace components. It was found that the original statistical analysis methods used in these four studies were either outdated or deficient in other aspects. Thus, the original POD data were reanalysed using the currently accepted approach in which maximum likelihood estimation was used to fit a log-normal cumulative distribution function to the POD hit/miss data as a function of crack size and statistical confidence levels were determined using the Q_2 likelihood ratio statistic. The data apply to wet fluorescent particle inspection using the continuous magnetisation method.

Analysis of the MPT POD studies showed a spread in performance between the organisations involved in the various trials. Following a meta-analysis, it was concluded that the a_{NDI} = 2.0 mm currently assumed by the Royal Australian Air Force is consistent with the *average* performance of MPT derived from the relevant available literature. The smallest a_{NDI} consistent with *most* implementations of MPT is 2.6 mm. On this basis, the results do not support a reduction in the standard limitation for MPT.

This page is intentionally blank

Authors

S.K. Burke
Maritime Platforms Division

Dr Stephen Burke is Head of the NDE Group in the Maritime Platforms Division at DSTO Melbourne. He completed a B.Sc. (Hons) at Monash University and subsequently a Ph.D. at Imperial College, London. Since joining DSTO, Dr Burke has been involved in the development of electromagnetic NDE techniques for structural integrity management of military aircraft, Navy ships and submarines. Dr Burke is a Fellow of the Australian Institute of Physics and a Chartered Physicist (Institute of Physics, UK). He is the author of over 40 publications in NDE and is the recipient of four TTCP Achievement Awards for international collaborative research in NDE.

R.J. Ditchburn
Maritime Platforms Division

Dr Robert Ditchburn is the Science Team Leader for Maritime NDE research in the Maritime Platforms Division at DSTO Melbourne. He received a B.App.Sc. (Hons) in applied physics and a Ph.D. from the University of Technology, Sydney. In 1993 Dr Ditchburn joined DSTO and has primarily been involved in research into NDE techniques for through-life support of Navy submarines and RAAF aircraft.

This page is intentionally blank

Contents

Abbreviations

AC	alternating current
ADF	Australian Defence Force
AFRL	Air Force Research Laboratory
AGARD	Advisory Group for Aerospace Research and Development
ASNT	American Society for Nondestructive Testing
ASTM	American Society for Testing and Materials
BHEC	bolt-hole eddy current
DC	direct current
DGTA	Director General Technical Airworthiness (ADF)
ECT	eddy-current testing
EDM	electrical discharge machining
ICON	Intercalibration of Offshore NDT
ISO	International Organization for Standardization
LCL	lower confidence limit
LPT	liquid penetrant testing
MLE	maximum likelihood estimation
MPI	magnetic particle inspection
MPT	magnetic particle testing
NDE	nondestructive evaluation
NDI	nondestructive inspection
NDT	nondestructive testing
NDTSL	Nondestructive Testing Standards Laboratory, DGTA ADF
NRC IAR	National Research Council Institute for Aerospace Research (Canada)
POD	probability of detection
RAAF	Royal Australian Air Force
RSRM	reusable solid rocket motors
RT	radiographic testing
UCL	University College London
UT	ultrasonic testing

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
t	thickness
Q_2	likelihood ratio statistic

1. Introduction

Knowledge of the reliability of defect detection in nondestructive inspection (NDI) of critical structures is essential to aircraft structural integrity management, whether during production or in service. For aircraft managed using a safety-by-inspection airworthiness philosophy, the interface between NDI and structural integrity management is via the quantity a_{NDI}, the smallest reliably detectable defect size for a particular inspection process for the structure in question. The value of a_{NDI}, together with information on crack growth rates and the critical crack size a_{crit}, is central to the analyses used to set inspection intervals to ensure aircraft safety.

The most robust method for determining a_{NDI} is through probability of detection (POD) trials, in which representative components containing representative in-service defects are inspected by a population of inspectors using a specific inspection procedure. Statistical analysis of the hit/miss data obtained in the trial is used to determine a POD curve as a function of defect size a. Statistical confidence limits are also calculated for the derived POD(a) curve. According to this approach, a_{NDI} is identified with $a_{90/95}$, the defect size for which the POD is 90% at a 95% statistical confidence level. Further detailed information on POD and NDI reliability, as well as the relationship between a_{NDI} and airworthiness standards, is given in a previous DSTO report by Harding and Hugo [1].

DSTO is conducting a series of literature reviews to examine the reliability of standard NDI methods used for Australian Defence Force (ADF) aircraft. The first review has been completed [2] and examines the reliability of liquid penetrant testing (LPT). The review identified twelve major studies on the reliability of post-emulsifiable LPT over the period 1968 to 2009. There was significant variability in the results of these studies and a meta-analysis was carried out to determine a range of key statistics. It was found that the value of a_{NDI} = 3 mm currently assumed by the Royal Australian Air Force for LPT is consistent with the average (median) performance demonstrated in the literature. The data did not support a reduction in the existing value of a_{NDI} assumed for LPT. It was noted that the largest a_{NDI} consistent with most (90%) of the LPT reliability studies was 5 mm.

The present report is concerned with magnetic particle testing (MPT) and is the second in the series. Magnetic particle testing is a mature nondestructive inspection method for the detection of surface-breaking or near-surface discontinuities in ferromagnetic steels and has been in use since the 1940s. Along with visual inspection and liquid penetrant testing, MPT is one of the most common methods for detecting surface-breaking cracks in metallic parts. If a part can be magnetised then MPT is usually preferred over LPT because parts can often be tested more quickly.

In conducting the literature review, we take into account the guidelines for evaluating published POD data given by Harding and Hugo [1]. In particular, these authors recommend that the following questions should be carefully considered when assessing published reliability studies:

1. How closely do the NDI technique, defect type and material used in the published POD trial match the application of interest, and how important are the differences?

2. Where were the system boundaries?
3. Who conducted the POD trial and did they have a specific agenda?
4. What has *not* been said?

In addition, we also take into account the extent to which the studies follow the recommendations for best practice in POD trial design and analysis [3,4].

A brief overview of the technique of MPT is presented in Section 2 and a summary of the available literature on the reliability of MPT is given in Section 3. Six key studies on MPT reliability for aerospace applications identified between 1968 and 2011 are described in more detail in Section 4 and are followed by a meta-analysis. In reviewing the results of these previous studies, it became apparent that a reanalysis of the original data using more modern statistical methods was needed. The results of this extensive reanalysis are presented in detail in the series of appendices which accompany this report. Conclusions and recommendations are presented in Section 6.

2. Overview of Magnetic Particle Testing

MPT is used in heavy engineering to inspect welds for surface-breaking discontinuities and, together with magnetic rubber inspection, also finds widespread use in aerospace applications for inspection of critical steel components. If a component or structure is fabricated from magnetic steel then MPT is generally the frontline inspection technique for surface-breaking defects, whether in production or in service.

A brief overview of MPT is given in this section. Further information is available in the ASNT Handbook on MPT [5], the Metals Handbook [6] and by consulting the relevant Australian and International Standards [7-9]

Like other magnetic inspection techniques, MPT relies on the leakage of magnetic flux in the vicinity of a discontinuity to enable detection of surface-breaking or, under favourable circumstances, near-surface defects. When a ferromagnetic part is magnetised, the magnetic lines of force (magnetic flux) are predominately contained within the part. The presence of a surface or subsurface discontinuity in the part will distort these lines of force and cause local magnetic flux to leak outside the surface. Fine ferromagnetic particles applied to the magnetised component are attracted to the area of flux leakage by the magnetic field gradient, creating an accumulation of particles (called an indication) that can be seen by eye. In this way, very fine discontinuities such as cracks become easily visible. Such indications occur on the surface of the part, directly above the location of the flaw, and the length and surface orientation of the flaw can usually be inferred directly.

The magnetic particles can be applied as a dry powder ('dry particle') or more commonly as a water- or hydrocarbon-based liquid suspension (known as magnetic ink) or spray aerosols. Dry-powder particles are much larger in size than the wet particles used in magnetic inks. MPT using the wet particle method is preferred for the detection of fine surface cracks. The

visibility of magnetic particle indications can be enhanced by using particles with either a fluorescent or coloured coating to contrast with the surface of the specimen. Viewing the results of a fluorescent magnetic particle test is carried out in a darkened environment using an ultraviolet light source.

Magnetisation of the test piece can be accomplished using a variety of direct and indirect methods tailored to specific component geometry and required magnetisation direction. The relevant Australian Standard [9] categorises these methods as follows:

1. *Current flow methods*. A large applied current is passed through the test piece to produce the magnetic field. Electrical contact is provided using either contact heads or current prods. The test piece is part of the electrical circuit.
2. *Magnetic flow methods*. Permanent magnets or AC/DC electromagnetic yokes (such as a Parker Contour probe) are used and the component is part of the magnetic circuit.
3. *Coil methods*. The applied magnetic field is provided directly by an insulated coil carrying a current. The test piece is either held in the vicinity of the coil or located inside the coil.
4. *Threading bar and threading cable methods*. For inspection of holes, tubes or other hollow components, the region of interest is magnetised by threading an insulated current-carrying conductor through the hollow component. In the threading bar (or central conductor) method, a single conducting rod is used. In the threading cable (or cable wrap) method, a flexible cable is threaded through and around the component multiple times.
5. *Induced current methods*. In this case, current is generated in the test piece by electromagnetic induction. The test piece acts in effect as the secondary winding of a transformer.

The magnetic fields or currents in these cases can be DC, AC, rectified AC, or pulsed, depending on the requirement. Thus, for MPT, inspection units can range from 240 V portable electromagnetic yokes to multipurpose fixed magnetic particle benches requiring a three-phase power supply.

The magnetisation of the component must be sufficient to ensure distinct indications but not so large as to produce noise or irrelevant particle accumulations which could mask any indications. Ideally, the required level of magnetisation is established on a case-by-case basis through systematic studies using specimens containing known discontinuities. Once established, the required magnetisation is controlled for each test.

To be detected, the discontinuity must also be favourably oriented to the direction of the magnetisation. For optimum detectability, the direction of magnetisation should be 90° to the line of the defect. However, it is possible to detect discontinuities which are up to ±45° to the direction of the magnetic field. If the orientation of expected discontinuities is not known then a part is normally tested at least twice by magnetising in at least two directions which are perpendicular to each other.

Magnetic particle testing can be performed using either the continuous technique, where particles are applied during the magnetising cycle, or the residual technique, in which the particles are applied after the magnetisation cycle has been completed and the component is in a remanent magnetic state. The continuous method offers the greatest sensitivity.

This literature review focuses on the two MPT methods most commonly used for ADF aircraft components:

- Fluorescent wet particle inspection, continuous method, using a stationary magnetic testing unit (magnetic particle bench) [10]

and, when the above method of magnetisation is not available or suitable,

- magnetic wet particle testing using portable magnetising yokes, continuous method, with either fluorescent particle aerosols or black ink aerosols [11].

The quoted RAAF limitation (or estimated a_{NDI})* for both methods is 2 mm (0.080 in.) surface length for surface-breaking discontinuities oriented at 90° to the magnetic field direction.

3. Literature Review

A survey of the published literature on the reliability of MPT was carried out with a broad initial coverage. Attention was paid not only to the reliability information that was presented but also to the strength of the underpinning evidence contained in the reviewed publication. The results of the survey are presented in this section in chronological order; first examining MPT for aerospace applications, then MPT for detection of surface defects in offshore welded structures, and concluding with a summary of reliability information embedded in aerospace standards.

3.1 Aerospace

It appears that the earliest reporting of POD results for MPT was in the late 1960s. A major study led by Packman from Lockheed-Georgia, first reported in 1968 [12] and subsequently in 1969 [13], examined the capabilities of four of the mainstream NDT techniques (MPT, LPT, ultrasonic testing and X-radiography) to detect and size laboratory-grown surface fatigue cracks in cylinders manufactured from 7075-T6511 aluminium alloy and 4330 vanadium modified steel. The magnetic particle inspections were carried out using the wet continuous method with fluorescent particles in both laboratory and production environments. Fractography was used to determine the geometry of each fatigue crack. The authors concluded that none of the four techniques could 'consistently' detect a defect smaller than 0.1 in. (2.54 mm) in length and that improvement to NDT sensitivity was required, especially for small cracks. For MPT on the 4330V steel cylinders, examination of the Packman results

* According to this terminology, the 'limitation' of the method is an estimate of a_{NDI} and its use for NDI reliability purposes is subject to a number of caveats [1]

indicates that the a_{90} value* lay in the range 3 – 6 mm for laboratory inspections and approximately 6 mm or greater for inspections in a production environment.

Three comprehensive studies on MPT reliability relevant to aerospace applications were subsequently carried out in the mid 1970s by: Southworth, Steele and Torelli from Boeing under contract to the US Air Force Materials Laboratory [14]; Rummel, Rathke, Todd, Tedrow and Mullen from Martin Marietta Aerospace under contract to NASA [15]; and Packman, Pearson, Owens and Young from Vanderbilt University and General Dynamics under contract to the US Air Force Office of Scientific Research [16,17]. These three studies examined a range of NDT techniques as well as MPT. In terms of the MPT aspects:

- Southworth et al. [14] evaluated the reliability of MPT for a range of defects and test-piece geometries in 4340M high-strength steel specimens. The defects included in the study were (i) hydrogen cracks, (ii) grinding cracks and (iii) artificial defects in the form of compressed electro-discharge machined (EDM) notches. Laboratory and production inspections were carried out using wet continuous fluorescent MPT. The laboratory inspections relied on magnetisation produced by a contour probe (magnetic yoke) whereas the production inspections used a magnetic particle bench.
- Rummel et al. [15] examined the reliability of MPT using a series of 4340 steel plate specimens containing fatigue cracks grown from EDM notches. Laboratory and production inspections were performed using the wet continuous fluorescent method. One complicating factor in this study was that two sets of inspections were performed. The first inspections were made with the specimens in the as-machined condition (i.e., after the starter EDM notches had been machined away) and the second inspections were made once the as-machined specimens had been etched and proof-loaded.
- Packman et al. [16,17] conducted an MPT POD study using laboratory-grown fatigue cracks grown from a weld or laser solidification spot in D6ac high-strength steel plates. The specific MPT inspection method was not described in the available documents. It could reasonably be assumed that the wet continuous fluorescent method was employed given (i) its widespread military aerospace use, and (ii) the wet continuous fluorescent MPT had already been used in an earlier MPT POD program led by Packman [12]. The results of the Packman study were highlighted almost 20 years later [18] in a review paper on failure analysis, NDI reliability and fracture mechanics, indicating a gap in the available literature on the reliability of MPT from the late 1970s to the early 1990s.

The results and implications of these three studies are considered in more detail in the analysis presented in Section 4.

For the record, it is worth noting that MPT was not included in the pioneering program to determine the reliability of NDT† in the USAF conducted in the 1970s, known colloquially as the 'Have cracks will travel' study [19].

* The terms POD and a_{90} were not in use at this time and Packman uses "sensitivity index" for POD.
† For the purposes of this report the terms nondestructive testing (NDT) and nondestructive inspection (NDI) are used interchangeably.

In a 1983 technical article published in *Materials Evaluation*, Hagemaier [20] stated that (aerospace) NDT programs had demonstrated a value of $a_{90/95}$ = 1.25 mm (0.05 in.) for MPT of fracture-critical components but neither details, references nor supporting evidence were given. A search of the available literature did not reveal any accessible publications to which Hagemaier may have been referring. The value quoted by Hagemaier is much lower than those reported in earlier published studies so that further details would have been extremely valuable. It is tempting to place some degree of weight on this result in view of Hagemaier's long-standing contributions in commercial aerospace NDT. However, in the absence of further information, this quoted result cannot be taken further and is excluded from the meta-analysis presented in Section 4.

In the mid 1990s, POD studies were conducted by the Thiokol Corporation (USA) on the reliability of MPT and eddy-current inspection for space shuttle reusable solid rocket motors (RSRM). The available literature on these studies appears to consist only of short conference papers [21,22]; detailed technical reports could not be located during the literature search. In the first paper, Hibbert et al. [21] describe a major POD study of MPT for RSRM ferromagnetic components involving 100 test specimens and 23 inspectors. Defects included fatigue cracks in plates and at holes, as well as stress corrosion cracking. The wet fluorescent MPT method was used at a range of magnetisation levels. As a result of this work, a_{NDI} values were determined for all areas of RSRM hardware—suggesting that different a_{NDI} values were assigned to different inspection geometries. Unfortunately, no actual a_{NDI} values could be found in the open literature. In a second conference paper, Hartman and Hibbert [22] report a value of $a_{90/95}$ = 4.4 mm for MPT of corner cracks at 25 mm diameter holes in D6ac steel. Again, in the absence of further published technical information on the Thiokol POD trials, this quoted selective result in a conference paper will have to be excluded from the meta-analysis in Section 4.

Two important studies into the reliability of a range of inspection techniques for aircraft gas turbine engine components were coordinated by the National Research Council of Canada Institute for Aerospace Research (NRC IAR) in the 1990s. These trials were distinguished by the use of real in-service defects in retired J85-CAN40 engine components rather than laboratory-grown fatigue cracks. The first study comprised a round-robin POD program involving six laboratories in four NATO countries. The MPT component of the study involved three laboratories and the POD results are reported in detail in two technical reports [23,24] with preliminary results reported in a conference paper [25]. The second study examined the performance of four Canadian organisations across a range of inspection techniques, however, only one organisation contributed MPT trial data. As with the first study, the results and analysis were reported in a detailed technical report [26] as well as further work in related conference papers [27-29].

As noted in the DSTO review on the reliability of LPT [2], the statistical analysis used by NRC IAR for the calculation of confidence intervals and hence $a_{90/95}$ is deficient and so the actual NRC IAR $a_{90/95}$ values will be disregarded in this literature review. A careful reanalysis of the NRC IAR data using currently accepted statistical methods to calculate the lower 95% confidence intervals was subsequently carried out by DSTO and is described in Section 4 and Appendix D.

The reliability of MPT is included in the *NDE Capabilities Data Handbook* compiled in 1997 by Rummel and Matzkanin [30]. In this Handbook, the authors aim to provide a single comprehensive source of reference data for the available POD trials at the time. The Handbook includes a reanalysis of the original data in which the maximum likelihood method (MLE) is used to fit a log-logistic curve to the POD data and obtain a_{90} values. The Handbook also gives the raw hit/miss and crack size data in electronic form, together with calculated curves. Lower 95% confidence intervals are also calculated and included in the data but $a_{90/95}$ values are not given explicitly. The authors chose not to plot the 95% confidence levels intentionally 'to emphasise the need for the user to independently validate NDE procedures that are used in critical designs'. The method used for calculating the confidence levels is not given.

The Handbook contains a reanalysis of three MPT POD trials mentioned previously, namely (i) the 1976 trials conducted by Rummel et al. [15] on 4340 steel plates, and (ii) the two NRC IAR studies on retired jet engine components [23,26]. As mentioned in the LPT study [2], the Handbook subdivides each study into the smallest possible subsets regardless of the final grouping that was used by the original researchers. For example, whilst Rummel et al. [15] pool the results from three inspectors to generate one POD curve, the Handbook presents POD curves for each inspector. Similarly, where NRC IAR combine the results of engine disc and spacer inspections to produce an overall POD curve for each organisation, the Handbook publishes separate POD curves for discs as well as spacers*. Thus the Handbook presents twelve POD curves compared with six in the original studies and reports values of a_{90} ranging from 1.0 mm to 17.56 mm. The latter value is due to the use of incorrect defect lengths in the Handbook reanalysis of the Rummel et al. [15] study and should be disregarded (see Appendix C for further details).

In a 1998 Netherlands National Aerospace Laboratory (NLR) internal report, Heida and Grooteman [31] quote a value a_{NDI} = 2.54 mm (0.10 in.) surface crack length for in-service MPT of F-16 airframe structure. This a_{NDI} value is attributed to the technical manual for the F-16A and F-16B aircraft [32]. These manuals were not available for this literature search. In the absence of published POD trials or other available evidence underpinning this value, it is excluded from further consideration in this report.

In more recent work, Piotrowski and Lee [33] presented results of a series of POD trials of wet continuous-method fluorescent MPT conducted in-house at Delta Air Lines. The available literature on these trials consists of a set of slides presented at the annual US Air Transportation Conference in 2008. Further information could not be located during the literature search. Three in-house studies were described:

- *MPT of threaded wheel bolts*. This trial was conducted using nine inspectors and a set of 46 wheel bolts. The bolts contained cracks at the threads, although the nature and source of the cracks is not described. The authors analysed the hit/miss data by fitting a log-logistic model and determined values of a_{90} corresponding to the nine inspectors. The results were not pooled. The a_{90} values for each inspector varied from 1.25 mm to

* The *NDE Capabilities Data Handbook* does not include the POD data for spacer inspections conducted by Organisation III which was published in the NDC IAR NATO study. For this data set, none of the cracks present (a < 1.2 mm) were detected.

6.24 mm, with an average a_{90} value of 3.2 mm. This average a_{90} value was also referred to as the $a_{90/95}$ value, which is possibly a typographic error. The authors noted that the a_{90} value was lower than the baseline value achieved in 2000.

- *MPT of flat 4130 steel plates containing holes*. A summary was given of previous '29/29' MPT POD tests conducted at Delta Airlines. The specimens comprised 4130 steel plates with laboratory-grown corner cracks in 12.7 mm diameter holes. In this type of test, a series of specimens containing 29 cracks of nominally the same size are inspected. If an inspector successfully detects 29 out of the 29 cracks, then a POD greater than 90% has been demonstrated with 95% confidence for that particular crack size, inspector and process. This type of test is typically used for procedure and personnel qualification and provides more limited information than a full POD trial [34]. It was reported that 'several operators qualified at the 0.05 in. (1.3 mm) level', indicating that these particular inspectors achieved $a_{90/95} < 1.3$ mm. Without further information on the performance of the entire inspector population involved in the tests, it is not possible to draw any further conclusions.
- *MPT of flat 17-4PH stainless steel plates*. A brief summary was given of an earlier MPT POD trial using laboratory-grown fatigue cracks in flat steel plates. Piotrowski and Lee [33] stated that Delta had demonstrated a 'capability at the 0.056 in. (1.42 mm) level' which 'exceeded industry standards'.

The main difficulty in further including these Delta Air Lines results in the present review is the absence of other published information beyond that which can be gleaned from this conference presentation, which by its nature, must be selective in its reporting due to length constraints. Further information on the POD trial for MPT of threaded bolts would have been particularly valuable because the Delta Air Line study appears to be the first of its kind which has been published.

Finally, in 2010, Hardwick and Moore [35] presented the results of a US Air Force Research Laboratory (AFRL) MPT trial using the same 46 threaded bolts used in the study by Delta Air Lines [33]. The AFRL study was conducted at five locations using a total of eleven inspectors. The AFRL presentation does not give numerical results, but it does state that the results 'did not meet expectations and were not consistent with the commonly observed USAF field capability'. The various cleaning methods used at the inspection locations is offered by the authors as one likely contributor to the lower than expected reliability. The available information on these trials again consists of the set of slides presented at the annual US Air Transportation Conference in 2010.

3.2 Offshore welded structures

Outside the aerospace domain, there have been a number of national and international projects undertaken to examine the reliability of MPT for detecting surface-breaking defects in welded structures, including above-water and underwater MPT of offshore platforms [36]. Four major offshore POD programs are summarised below.

- Initial studies on the reliability of underwater NDT for offshore structures were coordinated by University College London (UCL) in the late 1980s and early 1990s.

These studies considered the detection and sizing of fatigue cracks in offshore tubular welded joints using a range of NDT techniques, including underwater MPT [37]. Part of the motivation for these studies was to compare the effectiveness of underwater MPT with more advanced underwater electromagnetic techniques such as eddy-current NDT, AC Potential Difference (ACPD) and AC Field Measurement (ACFM). The cracks of concern in these studies are large by aerospace standards, with lengths over 100 mm. As an example, a_{90} = 50 mm surface length for low light underwater MPT on tubular joints [38] and in another case, $a_{90/95}$ = 7 mm crack depth [36].

- Following these initial UCL studies, an international collaborative project was initiated involving six organisations from the UK, France and Italy [38,39]. The project, entitled the *Intercalibration of Offshore NDT* (ICON), was also focused on the performance of underwater inspection of fatigue cracks in welds. Underwater MPT was compared with a range of other electromagnetic NDT techniques as well as ultrasonic inspection methods. The results are incorporated in a large database available to ICON users giving POD curves as a function of crack length and depth. Again, the defects investigated are large by aerospace standards: the critical range is 10 mm to 50 mm in length.
- In the mid 1990s UCL was also involved in the Topside Inspection Project. To date, we have been unable to obtain copies of the final report for Phases I and II. However, from the summaries of these reports given by Visser [36], this study was concerned detection and sizing of discontinuities in welded and unwelded offshore structural steel components. A range of NDT methods were compared, including MPT.
- Also, in the time frame 1984–1990, the Nortest NDT program [40] was carried out and involved 30 companies from the Nordic countries: Denmark; Finland; Norway and Sweden. The study involved a round-robin exercise to determine the reliability of detecting surface-breaking defects in ferritic and austenitic steels as well as aluminium alloys using MPT and LPT. There were 67 specimens containing 294 surface-breaking defects up to 75 mm in length available for magnetic particle inspection. The defects included porosity and slag inclusions as well as cracks. Inspections were conducted by both certified and non-certified inspectors. The results of this study were mainly reported in terms of defect depth (estimated using potential drop or eddy-current methods) rather than surface length. Destructive analysis was not carried out to verify defect hit/miss data by identifying additional cracks that may have been presented and remained undetected. The study was able to rank the effectiveness of MPT, LPT, X-radiography and eddy-current testing according to the defect size and type.

A good summary of these four projects is given by Visser [36].

These reliability studies are of little direct relevance to the present review of MPT for aerospace applications because of the differences in the metallurgy, surface condition, component geometry and residual stress state of welded structures compared with high-strength steel aerospace components. The practice of MPT also differs: for example, large welded components cannot be inspected using a particle bench. Finally, the structural certification and regulatory framework for the two sectors is different. With this in mind, and in the interests of relevance, deeper literature searches into the reliability of MPT for weld

inspection in other heavy engineering applications such as shipbuilding or pipeline inspection were not conducted.

3.3 Aerospace standards

A number of aerospace standards include, for engineering purposes, values of the minimum defect sizes that are assumed to be reliably detected by the NDI process. These defect sizes are default values to be used for structural integrity calculations in the absence of $a_{90/95}$ values obtained by a specific POD validation trial [1,2]. These values are generally conservative: smaller a_{NDI} values can be used if demonstrated by a POD trial.

Assumed in-service inspection initial flaw sizes are listed in Table XXXII of JSSG-2006 Guide to the Specification of Aircraft Structures [41]. These initial flaw sizes are the same for MPT as LPT (as well as for ultrasonic and eddy-current testing). The assumed values of a_{NDI} depend on the material thickness t and the nature of the crack (corner crack, surface crack or through-thickness crack). JSSG-2006 Table XXXII presents the following:

- a_{NDI} = 6.3 mm (0.25 in.) for corner cracks at holes in material with t > 6.3 mm
- a_{NDI} = 6.3 mm (0.25 in.) for through thickness cracks at holes in material with $t \leq 6.3$ mm
- a_{NDI} = 12.7 mm (0.5 in.) for surface cracks in material with t > 6.3 mm, semicircular crack shape
- a_{NDI} = 12.7 mm (0.5 in.) for through thickness cracks not at holes, material with $t \leq 6.3$ mm.

These values appear to be more conservative than the a_{NDI} values obtained from published MPT POD trials reviewed in Section 3.1 and later in Section 4.

The US Department of Defence Handbook for the Engine Structural Integrity Program, MIL-HDBK-1783B [42] assumes that manual NDT achieves $a_{90/95}$ = 1.8 mm (0.070 in.) surface length for surface cracks and $a_{90/95}$ = 0.89 mm (0.035 in.) for corner cracks. These values, taken from Table XVI of [42], are assumed to be the same for all manual inspection methods and so encompass MPT on ferromagnetic steel components. These values appear at the more optimistic end of the range of values described in Section 3.1 and Section 4, as also noted in the DSTO review of the reliability of LPT [2].

The two standards ISO 21347: 2005 [43] and NASA-STD-5009 [44] present separate $a_{90/95}$ values for MPT, LPT, eddy-current testing (ECT), radiographic testing (RT) and ultrasonic testing (UT). For wet fluorescent MPT, the two standards present the same table of values, for which the surface length of the minimum reliably detectable flaw size is:

- a_{NDI} = 6.4 mm for corner cracks at holes in material with t > 1.9 mm
- a_{NDI} = 6.4 mm for though cracks at holes in material with $t \leq 1.9$ mm
- a_{NDI} = 6.4 mm for surface cracks in material with t > 1.9 mm

- a_{NDI} = 9.6 mm for surface cracks in material with $t \leq 1.9$ mm
- a_{NDI} = 6.4 mm for through cracks not at holes, material with $t > 1.9$ mm.

Values are also presented in terms of the minimum detectable flaw depth rather than surface length. No details are given on how these values were determined, other than a note in [44] stating that they 'were derived from a limited set of specimens of simple geometry, and applying the crack sizes to complex geometries, other materials, material forms, material processes, and nonstandard NDE applications should be done with caution'. The values are broadly consistent with the values presented in JSSG-2006 but tend to be less conservative.

In terms of Australian documents, a Draft Civil Aviation Advisory Publication CAAP 42V-3(0) [45] states a procedure limitation of 2 mm surface length for (i) wet fluorescent MPT using a fixed particle bench and (ii) wet fluorescent MPT using a portable magnetic yoke. This is the first document written by CASA on the topic and was issued as a draft in 2006. Its current status is not known. These values are the same as the published RAAF limitations for the corresponding MPT general procedures [10,11] mentioned in Section 2.

The US Air Force have recently issued a structures bulletin EN-SB-08-012 [46] containing updated recommendations for a_{NDI} values to be used in calculations of reinspection intervals for structures managed through the USAF Aircraft Structural Integrity program. These a_{NDI} values are to be used when there are no supporting data available and were established on a consensus basis by the USAF NDI Capability Task Group, drawing on previous reliability studies, industry best practices and the experience of the group members. To date, there have been two documents issued, the first in 2009 and the second in 2011, and it is intended that the document will be expanded as new data become available. At present, the latest version of EN-SB-08-012 does not include a_{NDI} values for MPT.

4. Analysis

4.1 Selection of POD data from the literature

Having completed a review of the available literature, a judgement must be made as to which data to include in a meta-analysis* of the reliability of MPT. As discussed in Section 3, we exclude from further consideration the cases where a_{NDI} values are quoted but where little or no underpinning evidence has been presented or such information could not be located. Published a_{NDI} values embedded in aerospace standards are also excluded because, while these values are instructive, they are generally conservative default values. We also chose not to include the analysis presented in the *NDE Data Capabilities Handbook* but to rely on a DSTO reanalysis of the original studies described in the Handbook, in this way more statistically robust groupings of trial data could be used.

* In statistics, the term 'meta-analysis' describes the process of combining the results from related studies to obtain an overall conclusion which has greater statistical significance.

In order to refine the selection further, we consider '*How closely does the NDT technique, defect type and material used in the published POD trial match the application of interest, and how important are the differences?'* [1].

As the subject of the present report is aerospace applications, we exclude the studies conducted for MPT of welded structures in the maritime and offshore domains. The differences between MPT methods and practice between the two domains, as well as the differences in metallurgy and surface condition between machined thermo-mechanically processed high-strength steel aerospace components and large-scale welded structures are too significant for reliability studies on welded structures to be relevant to aerospace applications.

For the POD studies in the aerospace domain, the MPT method is common to all, i.e., continuous wet fluorescent MPT using a particle bench. Differences in the nature of the wet particles, carrier fluids, active magnetisation method or the type of particle bench are not expected to be important. The grade of high-strength aerospace steel used in the POD studies is also not expected to have a large effect on a_{NDI} provided the steel is adequately magnetised and the inspections are conducted on a clean, smooth, machined component, free from paint or other coatings. Differences due to the defect type and part geometry used in the POD trials will potentially influence a_{NDI} and are examined in more detail in Sections 4.2 – 4.5.

A more subtle question lies in the applicability of studies carried out as long ago as the late 1960s to current practice in 2012. Has the practice of MPT changed significantly over that time? In terms of the science and technology, MPT was introduced in the 1930s and wet fluorescent MPT was introduced in the 1940s [5], hence wet fluorescent MPT was certainly a mature technique by the 1960s. The methods of magnetisation were also very well established by that time. On this basis, it is argued that the changes in the science and technology of MPT since the late 1960s have been incremental and that the differences in equipment, magnetic field measurement or wet particle characteristics (while valuable) are not overwhelmingly influential. The more important difference is likely to be improvement in the training and accreditation of inspectors, and in strengthening the engineering management of the MPT inspection process. For the purposes of this analysis, the choice is made not to exclude any aerospace POD studies solely because of their age.

Six POD studies were selected for further evaluation and analysis on this basis:

(i) Packman et al. (1968): 4330V steel cylinders

(ii) Packman et al. (1976): D6ac steel plates

(iii) Southworth et al. (1975): 4340M steel cylinders, tubes and plates

(iv) Rummel et al. (1976): 4340 steel plates

(v) NRC IAR - NATO study (1994): Retired engine discs and spacers, AM355 steel

(vi) NRC IAR – Canadian study (1996): Retired engine discs, AM355 steel.

4.2 POD studies by Packman et al. (1968, 1976)

4.2.1 MPT of 4330V cylinders (1968)

According to the literature survey undertaken (Section 3.1), the study by Packman et al. [12] in 1968 was the first published trial to examine the reliability of MPT for detection and sizing of fatigue cracks for aerospace applications. Trials were conducted using laboratory-grown fatigue cracks in 4330 vanadium-modified steel cylinders with the crack dimensions established by destructive tests. Circumferential fatigue cracks were grown on the outer surface of annealed 4330V steel tubes and the cracked specimens then heat-treated to the correct temper condition before inspection. The inspections were carried out in both laboratory and production environments using the wet continuous fluorescent MPT method. Only a subset of the specimens was examined using production NDT. In this early study, hit/miss data were analysed in bins to provide point estimates of POD for a given crack size interval rather than the modern approach in which a POD curve is fitted to the raw hit/miss data. A later analysis [47] of the data contained in this report found that neither laboratory nor production MPT in these trials achieved a $a_{90/95}$ value for the range of flaw sizes investigated. In fact, none of the techniques studied achieved a $a_{90/95}$ value over the range of flaw sizes employed in this early study, suggesting deficiencies in the design of the trial.

A reanalysis of the results of this early POD trial was also attempted by DSTO using the currently accepted approach in which maximum likelihood estimation (MLE) is used to fit a log-normal cumulative distribution function to the hit/miss POD data as a function of defect size. Using the hit/miss data provided by Packman et al (Table XVII in the original report [12]), this approach returned values of a_{90} = 4.48 mm for laboratory inspections (66 cracks) and a_{90} = 6.90 mm for production inspections (45 cracks). Values for $a_{90/95}$ could not be computed from the data because the design of the trial did not include large enough cracks, rather than any deficiency in the analysis method. We do not attach any great weight to these computed a_{90} values because of a lack of confidence in the originally published data tabulations which contained unresolved inconsistencies in the numbers of small cracks between the tabulation of the hit/miss data (Table XVII in [12]), the summary tables of POD for given crack length intervals (Tables XIII and XV in [12]),) and the specimen schedule (Table II in [12]).

4.2.2 MPT of D6ac plates (1976)

In this later work, Packman et al. [16] examined the reliability of production MPT for detection of laboratory-grown fatigue cracks in D6ac high-strength steel plates. It was concluded that MPT was 'capable of high sensitivity in detection of flaws whose size was greater than 0.080 inches in length', which can be interpreted as a statement that a_{NDI} = 0.080 in. (2.0 mm).

The published report of the Packman et al. study [16] does not provide information on several key parameters, in particular the number of flaw sites, number of flawed test specimens and the false-call rate. Batches of specimens were cycled through the inspection process three times and the batches included 50% dummy specimens. The Packman study would not now be considered as best practice because of the low percentage of dummy specimens and small number of inspectors [3].

The Packman et al. [16] hit/miss data were reanalysed by DSTO, giving values of a_{90} = 2.32 mm and $a_{90/95}$ = 2.93 mm. These values are conservative estimates. Complete details of the reanalysis are presented in Appendix A.

4.3 POD studies by Southworth et al. (1975)

Southworth et al. [14] examined the reliability of MPT for inspection of high-strength 4340 steel for both laboratory and production inspections. The laboratory inspections relied on magnetising the specimens with a contour probe whereas the production inspections used a magnetic particle bench. The use of two methods of magnetisation in the trials is potentially useful because it may reveal a change in POD due to the magnetisation method. A range of specimen geometries and defect types were investigated.

The original Southworth et al. [14] hit/miss POD data were reanalysed by DSTO, as described in detail in Appendix B. The major observations are summarised and discussed below:

- *Cylinders, with and without fillet, containing an external flaw (compressed EDM notch)*

  There was no statistically significant difference in the reliability of the production and laboratory inspections (as shown for example in Figs B1 – B2 of Appendix B). For the laboratory inspections, a_{90} = 1.70 mm and $a_{90/95}$ = 2.05 mm, and for the production inspections, a_{90} = 1.60 mm and $a_{90/95}$ = 2.00 mm. This result indicates that the different methods of magnetisation (yoke compared with particle bench) for these trials had no measureable effect on the POD. Pooling the production and laboratory inspection data resulted in values of a_{90} = 1.66 mm and $a_{90/95}$ = 1.91 mm.

- *Hollow cylinders, with and without fillets, with flaws on the internal surface (compressed EDM notch)*

  In this case there was an extraordinary difference between the reliability of the production and laboratory inspections. While the laboratory inspections detected almost all of the defects (182 out of 184 inspection opportunities), the production inspections were successful in detecting the cracks in less than 50% of the inspection opportunities (48 out of 104 inspections). The laboratory inspections achieved similar reliability to that observed for externally flawed specimens. The production inspections failed to achieve 90% POD for any crack length.

  Southworth et al. [14] attributed this major difference to (i) reduced visual access when inspecting the inner surface of the tube using a particle bench compared with a contour probe and (ii) human factors related to the expected defect location. The authors indicate that the laboratory inspections were 'more carefully carried out' than the production inspections and that the performance of the production inspections improved when inspector experience increased. On this basis it is possible that the poor performance of the production inspections was due to an initial failure in the design of the human part of the NDT system (i.e., training and quality of the production inspection procedures for the trial).

  This large difference in reliability for MPT using a contour probe compared with a particle bench for detection of internal flaws contradicts the conclusion for MPT for external flaws. Thus, the documented evidence to support the use of the same a_{NDI} for

MPT using the different methods of magnetisation is equivocal and may depend on the location of the flaw.

- *Threaded cylinders with flaws in the threads (compressed EDM notch)*

  Little weight can be accorded to this element of the POD trials in view of the difficulties reported by Southworth et al. [14] in the preparation of the threaded cylinders.

- *Flat bars with hydrogen embrittlement cracks*

  A meaningful reanalysis of these data could not be performed because the pooled data consisted almost entirely of hits, so that the POD was always greater than 90% and essentially independent of crack length over the range of crack lengths used in the trial. Basically, not enough small hydrogen cracks were included in the design of the trial to perform the statistical analysis.

- *Flat bars with grinding cracks*

  The DSTO reanalysis of the data for MPT of grinding cracks was carried out separately to that of hydrogen embrittlement cracks rather than combining the two as had originally been done by Southworth et al. [14]. The two data sets were separated because the embrittlement cracks were isolated cracks whereas the grinding cracks consisted of a region of multiple cracks. The defect size a in the two cases is different: in one case it is the length of a single crack, in the other it is taken as the maximum dimension of the area of multiple cracking. As POD is determined as a function of defect size, it is not valid to combine data for the two. Reanalysis of the data for MPT of grinding cracks resulted in values of a_{90} = 6.25 mm and $a_{90/95}$ = 8.36 mm, where the size of the defect is the maximum dimension of the area of multiple cracking.

A noteworthy outcome from this study is that for externally flawed cylinders the two different methods of magnetisation (contour probe compared with particle bench) had a negligible effect on a_{90} and $a_{90/95}$ when detecting external flaws. This provides some evidence to support the use of the same limitation value in the two RAAF documents MPT/GEN/1 (Magnetic particle testing method for stationary magnetic testing units [10]) and MPT/GEN/2 (Magnetic particle testing using portable magnetising yokes [11]). The caveat is that significant differences in a_{90} and $a_{90/95}$ were observed when using the two different methods of magnetisation for MPT of internal flaws, suggesting that the flaw location may play a role.

The main drawback of the Southworth et al. [14] study was the use of compressed EDM notches to simulate defects rather than using cracks. Such defects are not truly representative of fatigue cracks: first, because the simulated defects had a triangular shape compared with semi-elliptical fatigue crack profiles; second, even though the slots are narrow, there is no possibility of crack-face closure and similar microstructural effects that would be present for in-service cracking. The effect on a_{NDI} compared with that of in-service cracking is not expected to be large, given that the compressed slots are narrow, surface-breaking planar defects interrupting the magnetisation of the specimen, but the magnitude of any influence on a_{NDI} is not known.

4.4 POD studies by Rummel et al. (1976)

Reanalysis of the trials conducted by Rummel et al. [15] for POD of fatigue cracks in 4340 steel plate gave rise to a number of challenges.

The first challenge, discussed in Appendix C, was that the electronic files containing the hit/miss data in the *NDE Capabilities Data Handbook* [30] contained incorrect crack length data. Some 37% of the tabulated crack lengths were incorrect, invalidating any analysis presented in the Handbook. It was therefore necessary to use the hit/miss data provided in the original 1976 report [15] in order to conduct the DSTO analysis.

A second challenge arose from the specimen manufacture. The fatigue cracks were grown from EDM starter notches which were then machined away when fatigue cycling had been completed. This resulted in flat plates containing cracks in known locations. In this 'as-machined' state, the presence of a cold-worked surface layer reduces crack detectability so the specimens were subsequently etched to remove the surface layer and subjected to a tensile proof load of 80% of the yield strength to ensure that the cracks were open to the surface. Inspections were carried out with the specimens in the as-machined state (which would lead to a conservative value of a_{NDI}) and also after etching and loads (which potentially leads to an optimistic value for a_{NDI}).

The final difficulty was that the test specimens contained a number of unplanned cracks, i.e., cracks that had not initiated from the EDM starter notches. The original authors did not include data for these unplanned cracks in their analysis because of difficulties in tracking and verification of the results. However, data for both the planned and unplanned cracks were included in a later analysis which appeared in the *NDE Capabilities Data Handbook*. DSTO performed a reanalysis using both data sets (planned cracks and all cracks) for the sake of completeness, but was guided by the decision of the original researchers, and gave preference to the results obtained for the data set containing the planned cracks.

Viewed in terms of currently accepted philosophies for the conduct of POD studies, this pioneering 1976 work by Rummel et al. [15] would now not be considered as best practice because of the limited number of inspectors (three) employed and the low number of defect-free specimens used (15 blank panels out of a total of 60 panels). Current best practice recommends POD trials include: (i) either 10 inspectors or 10% of the inspector population being studied (whichever is larger) and (ii) at least twice as many unflawed as flawed inspection sites [3]. The number of unintended flaws is also a complicating factor.

The full details of the reanalysis are given in Appendix C. For the set of planned cracks, values of a_{90} = 4.02 mm and $a_{90/95}$ = 5.14 mm were obtained for as-machined specimens and values of a_{90} = 2.01 mm and $a_{90/95}$ = 2.48 mm for specimens inspected after etching and proof loads. The improvement in a_{NDI} after etching and application of a proof load is concrete evidence of the importance of surface condition in MPT.

4.5 POD studies coordinated by NRC IAR (1994, 1996)

4.5.1 NATO AGARD Study (1994)

The MPT component of this round-robin reliability study involved inspections of retired aero-engine turbine discs and spacers by three NATO laboratories [23-25]. MPT was used to detect in-service low-cycle fatigue cracks originating from the disc and spacer boltholes. The crack geometry and dimensions were established by fractography after the holes had been broken open. Again, the results of the study were reanalysed by DSTO because of deficiencies (as discussed in [2]) in the calculation of the lower 95% confidence levels in the original reports. Adapting a later methodology employed by NRC IAR [26], the hit/miss data were pre-processed to exclude all data for very small cracks (a < 0.3 mm) in the DSTO reanalysis.

The recalculated a_{90} values are 2.39 mm, 1.46 mm and 1.54 mm for Organisations I, II and III respectively. The corresponding $a_{90/95}$ values are 2.96 mm, 1.81 mm and 2.25 mm. Full details of the reanalysis are given in Appendix D1.

4.5.2 Canadian Study (1996)

In this follow-on round-robin reliability trial [26], the MPT element was conducted by one Canadian organisation. Inspections were performed to detect in-service fatigue cracking in the boltholes of ten retired J95-CAN-40 engine discs, as in the preceding NATO AGARD trials [23,24]. Again, the results of the study were reanalysed by DSTO because of deficiencies in the calculation of the lower 95% confidence levels in the original reports. The raw hit/miss data were pre-processed to exclude all data for very small cracks. The reanalysis resulted in values of a_{90} = 2.44 mm and $a_{90/95}$ = 3.37 mm, somewhat larger but broadly consistent with the results of the NATO AGARD study (Section 4.5.1). The full details of the reanalysis are given in Appendix D2.

The defining feature of both the NATO AGARD trials and the follow-on Canadian study was the use of in-service cracks rather than laboratory-grown cracks or EDM slots. POD studies using in-service cracks are expected to achieve a higher level of fidelity than those which rely on artificial defects or laboratory-grown cracks.

A possible limitation of both the NATO AGARD and Canadian round-robin trials is that three distinct crack geometries (mid-bore cracks, corner cracks and through cracks) were grouped together in the MPT studies. Reflecting the component geometry, mid-bore and corner cracks dominate the crack population for smaller crack lengths (a < 1.5 mm) and through-thickness cracks comprise the bulk of the crack population for larger crack lengths (a > 1.5 mm). In later work, NRC IAR [27,28] show for automated eddy-current NDT that the POD curves for corner and mid-bore cracks are similar in this steel, but the POD curve for through-thickness cracks rise much more steeply. For MPT, which relies on flux leakage to produce crack indications but then human vision (rather than instruments) to detect the indication, it is not immediately obvious whether different POD curves apply.

4.6 Meta-analysis: Reliably detectable defect size for MPT

Having obtained the relevant data from the preceding studies, a meta-analysis can now be performed to estimate the reliably detectable defect size for MPT of aerospace components. A summary of a_{90} and $a_{90/95}$ values obtained from the reanalysis of the six selected POD trials (Sections 4.1–4.5) is presented in Table 1. On review, a number of these values will be excluded from the meta-analysis.

Of the twelve tabulated a_{90} values, we exclude four values on the basis of data integrity:

- The data from the early study by Packman et al. [12] on 4330V steel cylinders is not included because (i) the inconsistencies between the raw data and the tabulated summaries in [12,13] gave little confidence in performing a valid reanalysis and (ii) doubts remained on the maturity of the POD study, in particular, this was the only study for which a value for $a_{90/95}$ was not achieved (Section 4.2.1).
- The results obtained by including all cracks, rather than the planned cracks, in the reanalysis of the 1976 trials by Rummel et al. [15] (Section 4.4) are excluded. This choice is guided by the decision of the original researchers not to use data from unplanned cracks 'due to problems incurred in matching, verifying and correlating actual data to NDT observations'.

Of the eight remaining a_{90} values, two further values are excluded because the defect type is unrepresentative:

- The a_{90} value derived from the Southworth et al. [14] study is excluded because compressed EDM notches are not representative of fatigue cracks (Section 4.3).
- The a_{90} value obtained from the Rummel et al. [15] data for cracks in as-machined specimens is excluded because the presence of a cold-worked surface layer reduces crack detectability and machining over the top of existing cracks is not truly representative of the state of in-service surface-breaking fatigue cracks (Section 4.4).

Table 1 Summary of selected a_{90} and $a_{90/95}$ values calculated following DSTO reanalysis of the original published data

POD study	a_{90} (mm)	$a_{90/95}$ (mm)
Packman et al. (1969) 4330V steel cylinders—laboratory*	4.48	–
Packman et al. (1969) 4330V steel cylinders—production*	6.9	–
Southworth et al. (1975) Compressed EDM notch in 4340M steel—pooled	1.66	1.91
Packman et al. (1976) D6ac high-strength steel flat plates	2.32	2.93
Rummel et al. (1976) 4340 steel plate—as machined—planned cracks	4.02	5.14
Rummel et al. (1976) 4340 steel plate—as machined—all cracks	4.93	7.26
Rummel et al. (1976) 4340 plate—after etch and proof—planned cracks	2.01	2.48
Rummel et al. (1976) 4340 plate—after etch and proof—all cracks	4.57	6.21
Fahr et al. (1994) bolt holes in AM 355 engine discs and spacers—Org I†	2.39	2.96
Fahr et al. (1994) bolt holes in AM 355 engine discs and spacers—Org II†	1.46	1.81
Fahr et al. (1994) bolt holes in AM 355 engine discs and spacers—Org III†	1.54	2.25
Forsyth et al. (1996) bolt holes in AM 355 engine discs†	2.44	3.37

* $a_{90/95}$ values were not achieved for this POD trial.
† Pre-processing excluded all data (both hits and misses) for cracks < 0.3 mm in surface length

The results derived from the Rummel et al. [15] data for cracks after etching and proof load are included. It should be noted that such treatment presents a best case scenario for detection of fatigue cracks by MPT.

The final six a_{90} values derived from the POD literature for MPT after having made the exclusions on the basis of data integrity and defect type are plotted in Figure 1. There are identified shortcomings in all of the POD trials from which these data were derived. Nevertheless, when taken as a whole, these data can provide useful information on the expected reliability of MPT.

The median a_{90} value calculated directly from the sample data is 2.2 mm, with maximum and minimum values of 2.5 mm and 1.5 mm respectively, rounded to one decimal place. The variability in the a_{90} values (reflected in the sample variance) is smaller than reported in the related study of the available literature for LPT [2].

Harding and Hugo [2] present possible statistical measures which could be used to estimate an appropriate a_{NDI} given a range of a_{90} values derived from the literature, such as those presented in Figure 1. The approach assumes that the a_{90} values obtained from a reanalysis of the literature data follow a log-normal distribution and uses the maximum likelihood method to estimate the underlying parameters for the distribution. Details for the MPT data set are presented in Appendix E.

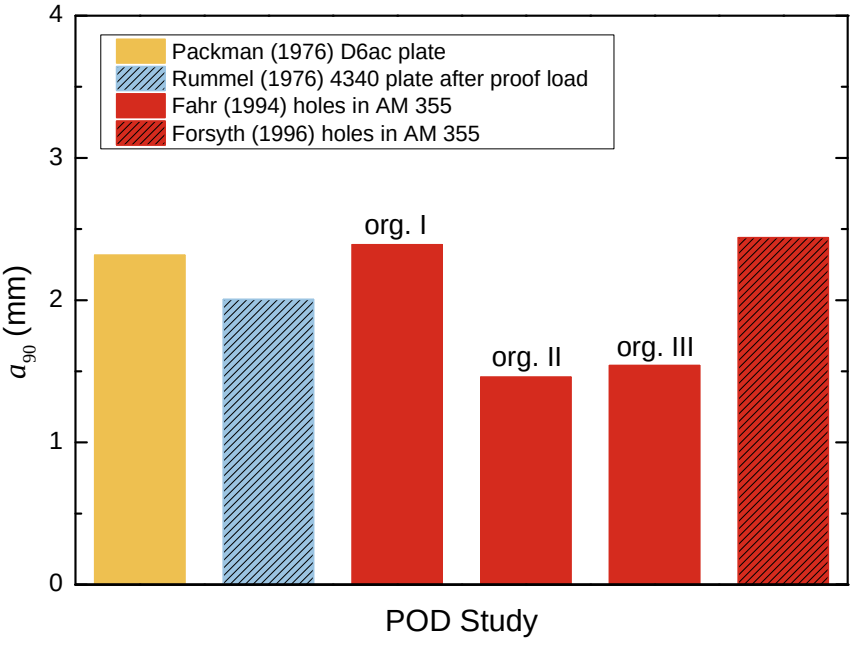

Figure 1 Summary of a_{90} values considered representative of MPT for fatigue-cracks in aerospace structures. The a_{90} values are calculated following a reanalysis of the original data given in the published literature.

Following this assumption, it can be shown that

- The best statistical estimate of the median a_{90} value for the data set is 2.0 mm (rounded to one decimal place) so that 50% of the a_{90} values are expected to be greater than 2.0 mm
- The predicted 90th percentile value for a_{90} is 2.6 mm, rounded to one decimal place, so that 90% of the a_{90} values are expected to be less than 2.6 mm

Thus, we conclude that the *average* (median) performance of MPT derived from the available literature corresponds to a_{90} = 2.0 mm, with 50% of the trials resulting in a larger a_{90} than this median value. The largest a_{90} consistent with *most* implementations of MPT is 2.6 mm, based on the 90th percentile value for the six selected a_{90} values.

Confidence limits can be placed on the median a_{90} to take into account the randomness associated with the small size of the sample set. According to the analysis presented in Appendix E, there is 90% statistical confidence that the true median a_{90} lies between 1.7 mm and 2.3 mm.

Similarly, confidence (or tolerance) limits can be placed on the estimated 90th percentile value for $a_{90.}$ As shown in Appendix E, the 90th percentile value for a_{90} lies between 2.2 mm and 3.4 mm with 90% statistical confidence. The upper tolerance limit (3.4 mm) is a highly conservative choice for a_{90} and could be unnecessarily pessimistic [2].

The values of $a_{90/95}$ obtained from these studies are not particularly useful for the meta-analysis because these results reflect as much the estimated statistical variability in the individual studies as the actual MPT performance. It is more reasonable to base the meta-analysis on the best estimates of a_{90} from each of the studies and to build in a level of conservatism through application of confidence levels to the resulting statistics. Nevertheless, it is possible to calculate an average value of $a_{90/95}$ for the entire data set by pooling the results from the six selected POD hit/miss data sets and performing an analysis on the combined hit/miss data. This is likely to be misleading if the trials are different in nature or if there is significant variability in the performance of the organisations participating in the trials [2]. For the sake of completeness, an analysis of the pooled data set for the six selected data sets was carried out, resulting in values of a_{90} = 2.2 mm and $a_{90/95}$ = 2.5 mm for the total pooled data. The corresponding notional POD curve is shown in Figure 2.

These statistical measures are compared in Table 2 together with the degree of confidence in whether each value could be used as a conservative *upper limit* on the performance of MPT for ADF aircraft [2].

These results do not support a reduction in the general limitation for MPT.

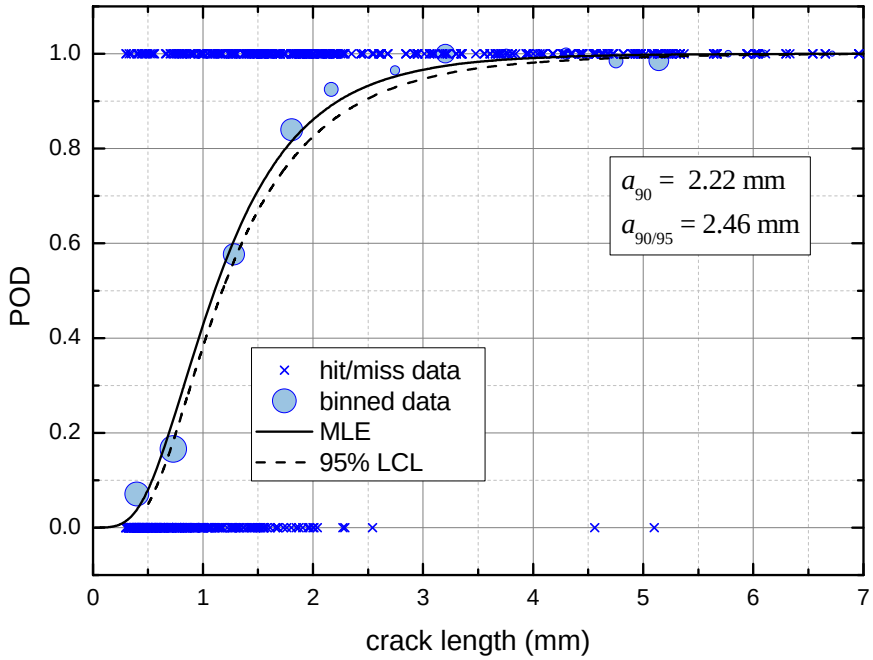

Figure 2 Notional POD curve for fluorescent MPT generated by pooling the hit/miss data from all six selected POD data sets. The total number of inspection results is n = 1329.

Table 2 summarises the uncertainty in the relevance of the studies to current ADF practice. For example, use of the median a_{90} as a measure of performance implies that there is a 50% risk that this value would not be achieved in practice. Similarly, use of the 90th percentile value for a_{90} implies that there is a 10% risk that this performance would not be reflected in practice. However, if current ADF practice is superior, the 90th percentile value for a_{90} would be an unduly pessimistic measure of performance.

Table 2 Summary of selected a_{90} and $a_{90/95}$ values calculated following DSTO reanalysis of the original published data

Statistic	Value (mm)	Degree of support* for use as an upper limit on MPT performance on ADF aircraft
Median a_{90}	2.0	Very low
Upper 90% confidence limit on the median a_{90}	2.3	Low
90th percentile value for a_{90}	2.6	Moderate
$a_{90/95}$ from pooled data	2.5	Moderate
Upper 90% confidence limit on the 90th percentile a_{90}	3.4	High

* Adopting the four point scale ranging from very low to high used by Harding and Hugo [2]

5. Discussion

A general observation from this survey is the paucity of recent, well conducted or fully documented reliability studies for MPT. Nevertheless, despite shortcomings in all of the POD trials examined, when taken as a whole it was possible to obtain useful information on the expected reliability of MPT for aerospace applications.

A number of statistical measures for the reliability of MPT were derived in the meta-analysis of the six published POD data sets and the level of risk in directly applying the results to Australian implementations was indicated. In the absence of $a_{90/95}$ values obtained from an Australian POD trial, these measures can be used to place the RAAF standard limitation in a broader context.

In the RAAF general procedures MPT/GEN/1 [10] and MPT/GEN/2 [11], the standard limitation of 2 mm crack surface length applies irrespective of crack geometry. Thus, the same limitation applies to corner cracks, surface cracks and through-thickness cracks irrespective of aspect ratio or depth. If the dominant detection mechanism is for the inspector to identify a line of fluorescent magnetic particles forming the crack indication in the presence of a noisy background (similar to LPT) then this is a reasonable assumption. However, if the crack geometry leads to smaller-than-expected magnetic leakage fields (such as for shallow cracks compared with semi-circular cracks) then this assumption is likely to be invalid because the indications will be weaker.

Similarly, the same standard limitations in general procedures MPT/GEN/1 and MPT/GEN/2, are assumed to apply irrespective of component geometry. This is a reasonable assumption for inspection of components with smooth machined surfaces, provided the component is adequately magnetised. There may be a case for a different procedure limitation to apply for (i) mid-bore cracks inside small diameter holes where visual access is more restricted, or (ii) cracks on the threaded surface of components where the threads may interrupt the formation and detection of small indications.

The only evidence found supporting the use of the same standard limitation for MPT for tests conducted using contour probes compared with particle benches relies on the one study (Section 4.3) for which contradictory results were reported depending on the flaw location.

In the DSTO reviews of the literature of POD for LPT and MPT, the RAAF standard limitations were independently found to be equal to the median a_{90} values obtained from analysis of the available literature. The close alignment of the standard limitations with the median in these cases is an interesting coincidence. The notable difference between the a_{90} values derived from the literature for MPT and LPT is the greater spread in performance for the various implementations of LPT. For MPT, the 90th percentile value is only 0.6 mm larger than the median whereas for LPT it is 2 mm larger.

The present review of the POD literature for aerospace MPT was significantly more time-consuming than initially anticipated because of the need for extensive checking and reanalysis of the original data. Hence, much of the length of the current report is devoted to

documentation of the methods, results and the judgements made in the DSTO reanalysis, contained in Appendices A – E. We note the need for care when using the *NDE Capabilities Data Handbook* [30]: the data contained in the Handbook for the Rummel et al. [15] studies of MPT, EC, UT, RT and LPT for 4340 steel specimens contain incorrect defect lengths and should not be used.

6. Conclusions

A review of the available literature on the reliability of magnetic particle testing identified some 20 references relevant to POD between 1968 and 2011. After critical examination, four published studies were considered both sufficiently well-documented and applicable to detection of fatigue cracks in high-strength steel aerospace components. As the statistical analysis methods used in these four studies were either outdated or deficient in other aspects, the original POD data were reanalysed using currently accepted techniques to give a series of six independent a_{90} and $a_{90/95}$ values. The data apply to wet fluorescent particle inspection using the continuous magnetisation method.

The MPT POD trials showed a spread of performance between the organisations involved in the various trials. Following a meta-analysis, it was concluded that the *average* performance of MPT derived from the available literature corresponds to a_{90} = 2.0 mm, with 50% of the trials resulting in a larger a_{90} than this median value. The RAAF standard limitation for MPT is 2.0 mm. The largest a_{90} consistent with *most* implementations of MPT is 2.6 mm, based on the 90^{th} percentile value for the six a_{90} values. On this basis, the results do not support a reduction in the standard limitation for MPT.

The literature review also identified a study in which it was found that the two different methods of magnetisation commonly used in MPT (contour probes and particle benches) had a negligible effect on a_{90} and $a_{90/95}$ for detecting external flaws. This provides some evidence, although equivocal, to support the use of the same limitation value in the two RAAF documents MPT/GEN/1 (Magnetic particle testing method for stationary magnetic testing units [10]) and MPT/GEN/2 (Magnetic particle testing using portable magnetising yokes [11]). In terms of reliability, the equivalence in performance of the two methods of magnetisation may depend on the flaw location.

A general observation from this survey is the paucity of recent reliability studies for MPT.

Acknowledgements

The authors wish to thank Dr Geoff Hugo for valuable discussions during the preparation of this report and Dr Cayt Harding for a critical review of the manuscript.

References

1. C.A. Harding and G.R. Hugo, 'Guidelines for interpretation of published data on probability of detection for nondestructive testing', Report No. DSTO-TR-2622, Defence Science and Technology Organisation, Australia (2011).
2. C.A. Harding and G.R. Hugo, 'Review of literature on probability of detection for liquid penetrant nondestructive testing', Report No. DSTO-TR-2623, Defence Science and Technology Organisation, Australia (2011).
3. J. Brausch, L. Butkus, D. Campbell, T. Mullis, and M. Paulk, 'Recommended processes and best practices for nondestructive inspection (NDI) of safety of flight structures', Report No. AFRL-RX-WP-TR-2008-4373, Air Force Research Laboratory, USA (2008).
4. *MIL-HDBK-1823A Nondestructive evaluation system reliability assessment* (Department of Defense, USA, 2009).
5. *Nondestructive Testing Handbook: Magnetic Testing*; *Vol. 8*, 3 ed., edited by D. G. Moore and P. O. Moore (American Society for Nondestructive Testing, Columbus 2008).
6. A. Lindgren, 'Magnetic particle inspection' in *ASM Handbook*; *Vol. 17 Nondestructive evaluation and quality control* (ASM International, 2001), pp. 89-122.
7. *ASTM E 1444-05 Standard practice for magnetic particle testing* (ASTM International, 2005).
8. *ASTM E 709-08 Standard guide for magnetic particle testing* (ASTM International, 2008).
9. *AS 1171 1998 Non-destructive testing – Magnetic particle testing of ferromagnetic products components and structures* (Standards Australia, 1998).
10. 'Non destructive testing general procedures—MPT/GEN/1—Magnetic particle testing method for stationary magnetic testing units', Report No. AAP 7002.043-36, Section 4, Chapter 1, Australian Defence Force, Australian Air Publication (1999).
11. 'Non destructive testing general procedures—MPT/GEN/2—Magnetic particle testing method using portable magnetising yokes', Report No. AAP 7002.043-36, Section 4, Chapter 3, Australian Defence Force, Australian Air Publication (1999).
12. P.F. Packman, H.S. Pearson, J.S. Owens, and G.B. Marchese, 'The applicability of a fracture mechanics–nondestructive testing design criterion', Report No. AFML-TR-98-32, Air Force Materials Laboratory, USA (1968).
13. P.F. Packman, H.S. Pearson, J.S. Owens, and G. Young, 'Definition of fatigue cracks through nondestructive testing', Journal of Materials, **4**, 666 (1969).
14. H.L. Southworth, N.W. Steele, and P.P. Torelli, 'Practical sensitivity limits of production nondestructive testing methods in aluminium and steel', Report No. AFML-TR-74-241, Air Force Materials Laboratory, USA (1975).
15. W.D. Rummel, R.A. Rathke, P.H. Todd, T.L. Tedrow, and S.J. Mullen, 'Detection of tightly closed flaws by nondestructive testing (NDT) methods in steel and titanium', Report No. NASA-CR-151098, National Aeronautics and Space Administration, USA (1976).
16. P.B. Packman, J.K. Malpani, and F.M. Wells, 'Probability of flaw detection for use in fracture control plans', Report No. AFOSR-TR-76-0290, Air Force Office of Scientific Research, USA (1976).
17. P.F. Packman, J.K. Malpani, F. Wells, and B.W.E. Yee, 'Reliability of defect detection in welded structures', Report No. AFSOR-TR-75-1603, Air Force Office of Scientific Research, USA (1975).
18. C.R. Morin, R.J. Shipley, and J.A. Wilkinson, 'Fractography, NDE, and fracture mechanics applications in failure analysis studies', Materials Characterization, **33**, 255 (1994).

19. W.H. Lewis, W.H. Sproat, B.D. Dodd, and J.M. Hamilton, 'Reliability of Nondestructive Inspections - Final Report', Report No. SA-ALC/MME 76-6-38-1, San Antonio Air Logistics Center (1978).
20. D.J. Hagemaier, 'A critical commentary on magnetic particle inspection', Materials Evaluation, **41**, 1063 (1983).
21. D.H. Hibbert, J.K. Hartman, and A.S. Allen, 'POD results of magnetic particle inspection of space shuttle RSRM hardware', 4th ASNT 1995 Spring Conference and Annual Research Symposium, Las Vegas, USA, (1995).
22. J.K. Hartman and D.H. Hibbert, 'Probability of detection results of eddy current inspection for ferromagnetic space shuttle solid rocket boosters', 4th ASNT 1995 Spring Conference and Annual Research Symposium, Las Vegas, USA, pp. 63-65 (1995).
23. A. Fahr, D. Forsyth, M. Bullock, W. Wallace, A. Ankara, L. Kompotiatis, and H.F.N. Goncalo, 'POD assessment of NDI procedures using a round robin test', Report No. AGARD-R-809, North Atlantic Treaty Organization (1995).
24. A. Fahr, D. Forsyth, M. Bullock, and W. Wallace, 'NDI techniques for damage tolerance-based life prediction of aero-engine turbine disks', Report No. LTR-ST-1961, Institute for Aerospace Research, Canada (1994).
25. A. Fahr, D.S. Forsyth, M. Bullock, and W. Wallace, 'POD assessment of NDE procedures—results of a round robin test', in *Review of Progress in Quantitative Nondestructive Evaluation*, edited by D. O. Thompson and D. E. Chimenti (Plenum Press, New York, 1995) Vol. 14, pp. 2391-2398.
26. D.S. Forsyth and A. Fahr, 'The sensitivity and reliability of NDI techniques for gas turbine component inspection and life prediction', Report No. LTR-ST-2055, Institute for Aerospace Research, Canada (1996).
27. A. Fahr and D.S. Forsyth, 'POD measurement using actual components', Proc. SPIE, **3397**, 194 (1998).
28. A. Fahr and D.S. Forsyth, 'POD assessment using real aircraft engine components', in *Review of Progress in Quantitative Nondestructive Evaluation*, edited by D. O. Thompson and D. E. Chimenti (Plenum Press, New York, 1998) Vol. 17, pp. 2005-2012.
29. D.S. Forsyth, A. Marincak, and J.P. Komorowski, 'Edge of Light: A new enhanced optical NDE technique', Proc. SPIE, **2945**, 178 (1996).
30. W.D. Rummel and G.A. Matzkanin, 'Nondestructive evaluation (NDE) capabilities data handbook', Report No. NTIAC-DB-97-02, Nondestructive Testing Information Analysis Center, Austin, USA (1997).
31. J.H. Heida and F.P. Grooteman, 'Airframe inspection reliability using field inspection data', Report No. NLR-TP-98144, National Aerospace Laboratory NLR, Netherlands (1998).
32. 'Nondestructive inspection; USAF/EPAF series F-16A and F-16B aircraft, Technical Manual T.O. 1F-16A-36, Change 38', Lockheed Martin Corporation (1997).
33. D. Piotrowski and J. Lee, 'Comparative assessment of probability of detection studies for FPI and MPI at Delta', 51st Annual Air Transportation Association (ATA) NDT Forum, Seattle, WA, USA, (2008).
34. C.A. Harding, 'Methods for Assessment of Probability of Detection for Nondestructive Inspections', PhD Thesis, The University of Melbourne, 2008.
35. J. Hardwick and C. Moore, 'US Air Force magnetic particle inspection field capability study', 53rd Annual Air Transportation Association (ATA) NDT Forum, Albuquerque, NM, USA, (2010).

36. W. Visser, 'POD/POS curves for non-destructive examination', Report No. 2000/018, Offshore Technology Report, Health and Safety Executive, UK (2002).
37. W.D. Dover and J.R. Rudlin, 'Underwater inspection reliability for offshore structures', Ship Structures Symposium '93, Arlington, Virginia, USA, pp. I-1-I-7 (1993).
38. J.R. Rudlin, 'Investigation of underwater MPI procedures in the ICON project', Insight: Non-Destructive Testing and Condition Monitoring, **38**, 415 (1996).
39. J.R. Rudlin and W.D. Dover, 'The ICON Project - data for underwater inspection', Insight: Non-Destructive Testing and Condition Monitoring, **38**, 412 (1996).
40. P. Kauppinen and J. Sillanpää, 'Reliability of surface inspection techniques', International Journal of Pressure Vessels and Piping, **54**, 523 (1991).
41. *JSSG-2006 Joint service specification guide. Aircraft structures* (Department of Defense, USA, 1998).
42. *MIL-HDBK-1783B Engine structural integrity program (ENSIP)* (Department of Defense, USA, 2004).
43. *ISO 21347:2005 Space systems – Fracture and damage control* (International Organization for Standardization, Geneva, 2005).
44. *NASA-STD-5009 Nondestructive evaluation requirements for fracture-critical metallic components* (National Aeronautics and Space Administration, Washington DC, USA, 2008).
45. *Draft CAAP 42V-3(0) Magnetic particle inspection – Use and implentation of ASTM-E-1444* (Civil Aviation Safety Authority Australia, 2006).
46. *EN-SB-08-012, Revision B Nondestructive inspection capability guidelines for United States Air Force aircraft structures* (ASC/EN WPAFB USAF Dayton Ohio, 2011).
47. B.G.W. Yee, F.H. Chang, J.C. Couchman, and G.H. Lemon, 'Assessment of NDE reliability data', Report No. NASA-CR-134991, National Aeronautics and Space Administration, USA (1975).
48. C.A. Harding and G.R. Hugo, 'Statistical analysis of probability of detection hit/miss data for small data sets ', in *Review of Progress in Quantitative Nondestructive Evaluation*, edited by D. O. Thompson and D. E. Chimenti (Plenum, New York, 2003) Vol. 22, pp. 1838-1844.
49. *E2862-12 Standard practice for probability of detection analysis for hit/miss data* (ASTM International, 2012).
50. P.F. Packman, 'Fracture toughness and NDT requirements for aircraft design', Nondestructive Testing, **6**, 314 (1973).
51. *MIL-STD-1949A Inspection-Magnetic particle* (Department of Defense, USA, 1989).

Appendix A: Reanalysis of Packman et al. (1976) data

Packman et al. [16] carried out a study of the reliability of MPT for detection of fatigue cracks in D6ac high-strength steel plates. The results of the study were published in the form of either cumulative POD or POD at a given confidence level for a range of flaw sizes. Binomial statistics were used within each flaw size range to estimate the confidence levels.

In this Appendix, the results presented by Packman et al. [16] are re-examined using a more modern approach to POD analysis through the POD-Q2 software (version 2.0.1) developed by DSTO. The software uses maximum likelihood estimation (MLE) to fit a log-normal cumulative distribution function to hit/miss POD data as a function of defect size. The input data are the set of ordered pairs of the form (defect size, hit/miss) where the defect size is the size of the defect inspected and the hit/miss parameter is the result of the inspection (a hit = 1 or a miss = 0 for that particular inspection). The one-sided lower 95% confidence level on the whole POD curve is calculated using the likelihood ratio statistic Q_2 [34,48]. The use of the Q_2 statistic has the benefit of an increased range of validity and can be used for small data sets (in some cases as few as 50 hit/miss observations). A similar approach to the analysis of hit-miss POD data is adopted in the most recent revision of MIL-HDBK-1823 [4] and more recently in ASTM E2862–12 [49].

A complete reanalysis of the Packman et al. [16] data is not possible because the original results are only provided in 'binned' form, i.e., the hit/miss data are only presented as totals over a range of crack lengths, as shown in Table A1, rather than as individual hits or misses for a specific crack length. To proceed with the analysis, the conservative assumption was made that the crack length corresponding to the binned hit/miss data was the largest crack length in the range. For example, referring to Table A1, it is assumed that the two hits and three misses in the flaw size range 0 – 1.0 mm occur for a crack length of 1.0 mm, the six hits and one miss in the flaw size range 1.0 – 1.5 mm occur for a crack length of 1.5 mm etc.

Table A1 Raw POD data extracted from Table VI of the Packman et al. 1976 reliability study of production MPT for D6ac steel plates [16]. The authors provide the hit/miss data arranged in bins according to the surface crack length ('flaw size range'). The final column showing the ratio of hits to observations has been added for convenience.

Flaw size range		Number of	Number	Number of hits /
inch	mm*	observations	of misses	Number of observations
0.00 – 0.04	0 – 1	5	3	0.400
0.04 – 0.06	1 – 1.5	7	1	0.857
0.06 – 0.08	1.5 – 2	26	6	0.769
0.08 – 0.125	2 – 3.2	86	0	1.000
0.125 – 0.20	3.2 – 5.1	91	2	0.978

* The Packman et al. data are presented in inches and are converted here to mm

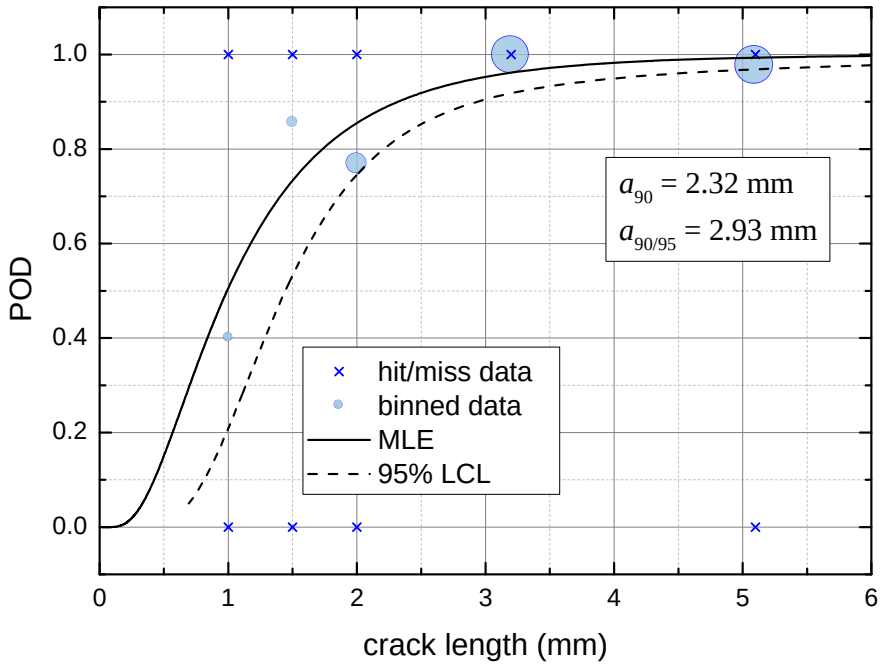

Figure A1 Reanalysis of Packman et al. (1976) data for production MPT of fatigue cracks in D6ac steel plate

The results of the DSTO analysis are shown in Figure A1, giving values of a_{90} = 2.3 mm and $a_{90/95}$ = 2.9 mm. This is consistent with the observation by Packman et al. [16] that the POD in this study exceeds 96% with 95% confidence for crack lengths in the flaw size range 2.0 – 3.2 mm. It is also consistent with an earlier statement by Packman that the POD for fluorescent MPT of fatigue cracks in D6ac steel exceed 90% for crack lengths greater than 2.5 mm [50]. As an indication of the sensitivity of the results to the assignment of the crack length within the flaw size range, assigning the crack length to midpoint of the range results in calculated values of a_{90} = 2.1 mm and $a_{90/95}$ = 2.85 mm.

In an interim 1975 report on the same study, Packman et al. [17] present the binned hit/miss data for MPT of the D6ac plates using slightly different flaw size ranges to those adopted in the 1976 report. A similar reanalysis of these earlier data, again assigning the crack length to the largest crack length in the range, results in calculated values of a_{90} = 2.1 mm and $a_{90/95}$ = 2.8 mm.

Appendix B: Reanalysis of Southworth et al. (1975) data

B.1 Introduction

Southworth et al. [14] report the results of a comprehensive POD study of defects in high-strength 4340M steel. The study examined the performance of several NDT techniques, including fluorescent MPT using the wet continuous method. This work was performed by Boeing Commercial Aircraft Company under contract to the USAF Materials Laboratory. The inspections were carried out in accordance with the Boeing process specifications and inspection procedures current at the time.

The study examined three classes of defects: (i) compressed EDM notches, (ii) hydrogen cracks and (iii) grinding cracks. The compressed EDM notches were wholly artificial defects produced by inserting an EDM slot into a circumferential groove in a specimen blank, applying axial compression to the specimen so that the EDM slot was closed by plastic deformation and then machining away the groove. The specimen was then heat-treated and ground. The resulting discontinuity had a rounded triangular shape with a surface-length to depth ratio of ≈ 2 with a typical opening width of 15 μm. Such artificial defects are not necessarily representative of tight fatigue cracks. Hydrogen embrittlement cracks were produced by electrolytic hydrogen charging of selected areas during the application of a bending stress. Grinding cracks were generated by charging a selected region of the specimen with hydrogen and then grinding the surface. This procedure resulted in an area of multiple grinding cracks. Crack depths were not reported.

Inspections were carried out for the following specimen geometries:

- Cylinder containing an external flaw (compressed EDM notch)
- Cylinder with fillet containing an external flaw (compressed EDM notch)
- Hollow cylinder with flaws on the internal surface (compressed EDM notch)
- Hollow cylinder with flaws in an internal fillet (compressed EDM notch)
- Threaded cylinder with flaws in the threads (compressed EDM notch)
- Flat bars with hydrogen embrittlement cracks
- Flat bars with grinding cracks

The study involved 30 inspectors from several facilities. Portable magnetic yokes (DA-200 Parker contour probes) were used almost exclusively for the series of 'laboratory inspections' whereas a stationary particle bench was used for the series of 'production inspections'. A summary of the test specimens used is given in Table B1. The completed inspection results can be found in the original report [14].

The main limitation of the study by modern standards was that a relatively small number of specimens were cycled through a large number of inspectors. This was noted by the authors who recommended in future studies '[to] perform fewer tests on a larger number of specimens', in part to avoid 'local wear patterns that signal flaw locations' to the inspectors, but also to draw on a larger flaw population. The false-call rate was reported as 13%, a significant fraction were due to irrelevant material discontinuities.

Table B1 Summary of test specimens and flaws used in the Southworth et al. MPT POD study for 4340M steel (Table 1 of reference [14]). The number of dummy unflawed specimens of each type is given in parentheses.

Configuration	Number of specimens	Number of flaws	Flaw size range (mm)
Cylinder, straight – external flaws	14 (9)	28	0.76 – 19.0
Cylinder, fillet – external flaws	15 (7)	30	0.50 – 12.7
Tube, straight – internal flaws	13 (6)	23	1.5 – 14.5
Tube, fillet – internal flaws	14 (6)	22	1.8 – 11.7
Cylinder, threads – external flaws	12 (7)	29	1.0 – 16.5
Flat bars – hydrogen cracks	7 (8)	10	1.3 – 13.7
Flat bars – grinding cracks	7 (?)	12	Multiple cracks

Southworth et al. [14] analysed the results from the POD trials by arranging the hit/miss data into bins according to the defect surface length and then using binomial statistics to determine the POD and the corresponding confidence level within each bin (surface length interval). This analysis method is now outdated and so a reanalysis was attempted. The reanalysis met with mixed success, as described below.

B.2 Results of reanalysis

As with the Packman et al. study [16] described in Appendix A, a complete reanalysis of the Southworth et al. data was not possible because the original data are only provided in binned form rather than as individual hits or misses for a specific crack length. Thus, as in Appendix A, a conservative approach was taken and the crack length corresponding to the binned hit/miss data was assumed to be the largest crack length in the range. Following Southworth et al. [14], the data for solid cylinders (with or without fillet) were combined for the purposes of analysis, as were the data for hollow cylinders (with and without fillet). However, in contrast to Southworth et al. [14], we examined the results for hydrogen cracks and grinding cracks separately. This is because the grinding cracks consisted of multiple defects distributed over an area of the specimen whereas the hydrogen cracks appear as single linear features so that the two are not the same class of defect. The effect of false calls was not included in our reanalysis.

B.2.1 Cylinders with external flaws

The results for compressed EDM notches on cylinders are shown in Figure B1 for production inspections and Figure B2 for laboratory inspections. There was no significant difference between the a_{90} or $a_{90/95}$ values for these two inspections. It is noteworthy that the laboratory inspections relied on the use of contour probes whereas production inspections used magnetic particle benches.

Analysis of the combined production and laboratory inspection results (not plotted here) gives values of a_{90} = 1.66 mm and $a_{90/95}$ = 1.91 mm. Pooling the data in this way is valid for

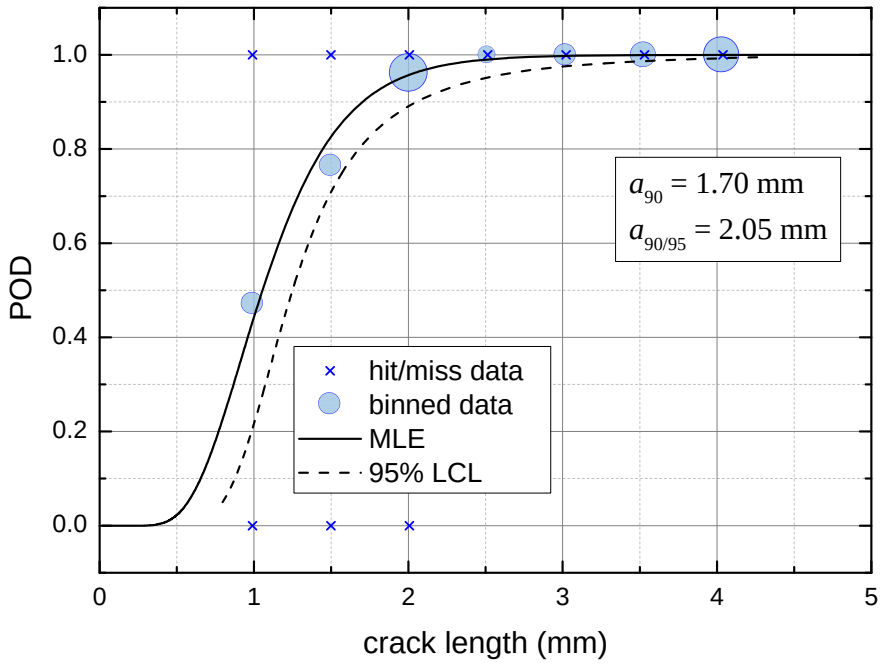

Figure B1 Reanalysis of Southworth et al. (1975) data for production fluorescent MPT of artificial defects in steel specimens (Table 8 in [14]).

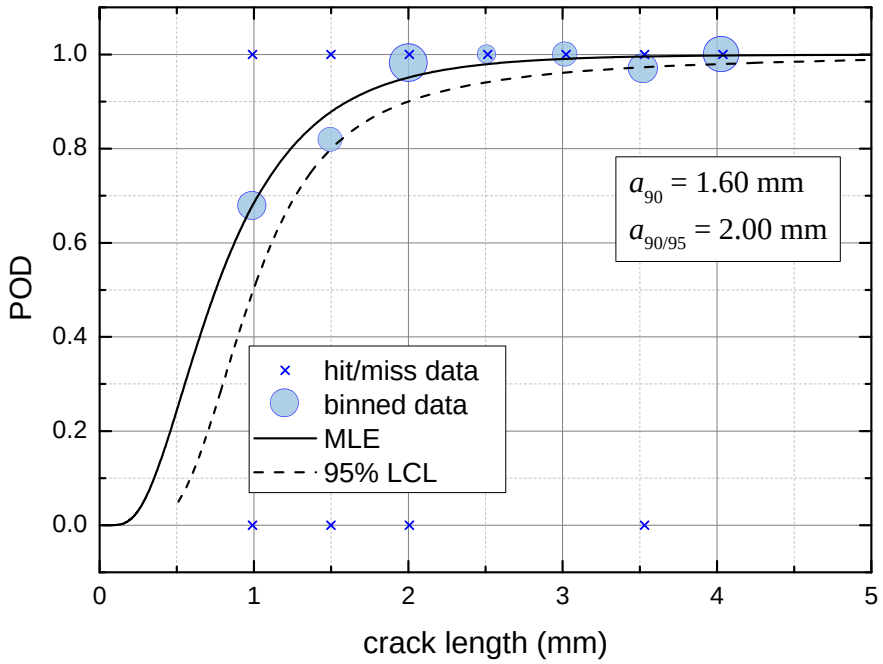

Figure B2 Reanalysis of Southworth et al. (1975) data for laboratory fluorescent MPT of artificial defects in steel specimens (Table 8 in [14]).

such similar data sets* and provides a larger sample size which reduces the statistical uncertainty in the estimate of a_{90}.

B.2.2 Hollow cylinders with internal flaws

In contrast to the inspection results for the cylinders containing external flaws, there was a dramatic difference between production inspections and laboratory inspections for the hollow cylinders containing flaws on the internal surface. While the laboratory inspections detected almost all of the defects (182 out of 184 inspection opportunities) with a reliability similar to that for the externally flawed cylinders, the production inspections failed to detect flaws in over 50% of inspections (48 hits out of 104 inspection opportunities). The original 1975 analysis indicates that $a_{90/95}$ lies between 1.5 mm and 2.0 mm for laboratory inspections, similar to that achieved for external flaws (Section B.2.1). The production inspections failed to achieve 90% POD for any crack length. Southworth et al. [14] attribute this difference to (i) reduced visual access when inspecting the inner surface of a tube in the production inspections and (ii) human factors related to expectations of likely defect locations. Reading between the lines, we speculate that the production inspections may have suffered through the lack of a detailed inspection procedure specific to these flaws.

A reanalysis of the laboratory inspection data using MLE curves was not possible because the data consisted almost entirely of hits for the range of defect sizes used in the trial.

B.2.3 Threaded cylinder with flaws in the threads

Difficulties were reported in the preparation of threaded cylinders containing compressed EDM notches for flaws, which were reflected in the inspection results. These difficulties included closely spaced defects (so that indications overlapped) and variability in defect location on the thread (some at the crest of the thread, some at the root of the thread). Southworth [14] noted that 'most of the [inspection] methods [were] penalised [by] peculiarities of the arrangement and distribution of flaws within this specimen design…' and that 'extensive inspections were not performed and more emphasis was placed on the other externally flawed steel specimens'. The 'limited results obtained…should not be directly compared with the capabilities demonstrated on other designs.' For the record, the DSTO reanalysis of these data gave values of a_{90} = 3.59 mm and $a_{90/95}$ = 7.13 mm for production inspections. Little weight should be placed on these numbers in view of the limitations described by Southworth for this data set.

B.2.4 Flat bars with hydrogen embrittlement cracks

The results of laboratory and production inspections of bars containing hydrogen embrittlement cracks were combined for the purposes of analysis. A meaningful reanalysis could not be performed because the experimental POD values were all greater than 90% and essentially independent of crack length over the entire range of crack lengths (Figure B3). Basically, the design of the Southworth et al. [14] trials for hydrogen embrittlement cracks did not include enough small cracks to allow a determination of a_{90}. The significance of these

* The upper and lower 95% confidence bands for the two data sets overlap over the entire range of defect sizes indicating that there is no statistically significant difference between the two data sets.

results should be treated with caution because a small number of flaws (12) were inspected by a large number of inspectors. Such a small sample size is not considered best practice in POD trials.

B.2.5 Flat bars with grinding cracks

As mentioned previously (Section B.2), the reanalysis for grinding cracks was carried out separately to that of hydrogen cracks rather than combining the data as Southworth et al. [14] had done originally. This separation was made because (i) the grinding cracks occurred as regions of multiple cracking rather than as single isolated cracks and (ii) the grinding cracks were likely to be a different depth than the equivalent hydrogen cracks for the same surface crack length. The reanalysis of the grinding crack trial data resulted in values of a_{90} = 6.25 mm and $a_{90/95}$ = 8.36 mm. It is important to note that in these results the defect size a is the maximum dimension of the area covered by the multiple grinding cracks so that the individual crack dimensions are smaller than a. It is therefore invalid to compare the a_{NDI} values above for grinding cracks to a_{NDI} values obtained for single isolated cracks, and underlines the difficulty in finding appropriate metrics for detection of multiple defects.

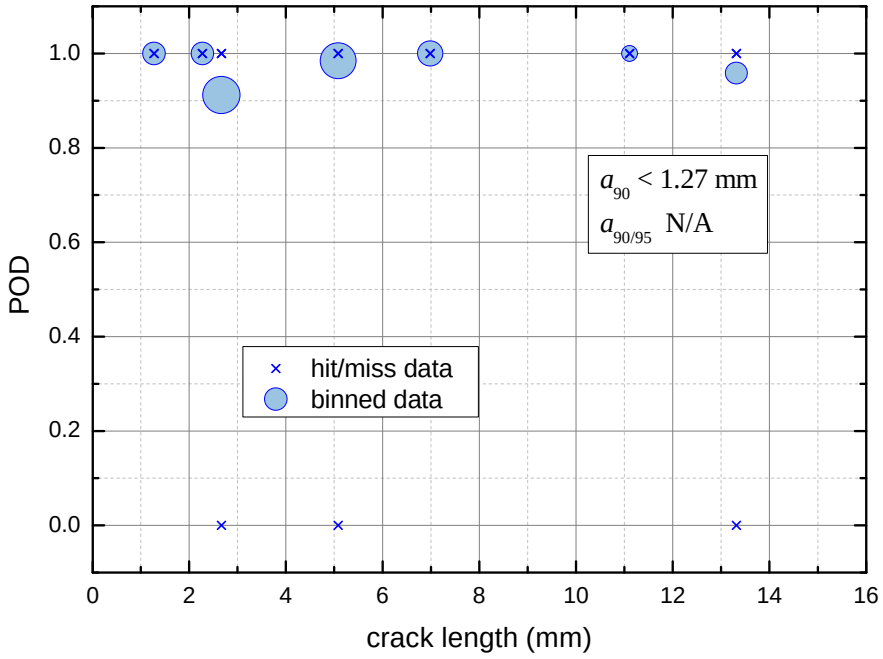

Figure B3 Reanalysis of Southworth et al. (1975) data for MPT of hydrogen cracks in steel specimens (Table 36 in [14]). The data for production and laboratory inspections are combined. A useful analysis of these data was not possible because there were no data points with POD values below 90% and so the MLE curve could not be fitted in a meaningful way.

Appendix C: Reanalysis of Rummel et al. (1976) data

As part of a larger study performed under contract to NASA in 1976, Rummel et al. [15] examined the detectability of tight fatigue cracks in 4340 grade steel plate specimens. In the original 1976 report, the POD analysis was performed using an overlapping sampling technique. The data were subsequently reanalysed in the *NDE Capabilities Data Handbook* [30] by fitting a log-logistic model to the original hit/miss data. Values for a_{90} were reported for fluorescent MPT but not values for $a_{90/95}$.

During our review of the analysis of the Rummel et al. [15] MPT data as presented in the *NDE Capabilities Data Handbook*, it became apparent that the Handbook had used incorrect crack dimensions in calculating a_{90}, leading to invalid results. By checking the crack dimensions in the original report (Table 9 in [15]) against those in provided electronically in the CD accompanying the Handbook (Excel® files B1001AL, B1001AD, B1001BL, B1001CL, B1001CD, B1003AL, B1003AD etc. in folders B-MT1L and B-MT1D [30]), it was found that a significant fraction* (up to 37%) of the tabulated crack lengths were incorrect. The crack depths were also in error. While some of the errors appeared to be simple transcription errors between the original report and the Handbook, the majority appear to have arisen from an inadvertent use of the Excel® autocomplete function for crack numbers 4 to 66. The error is also propagated in the Handbook analyses of the related Rummel et al. [15] studies of EC, UT, RT and LPT methods for the same 4340 grade steel specimens.

In view of these errors, the original Rummel et al. [15] hit/miss data as a function of crack length were reanalysed. As in the original work by Rummel et al. [15], the hit/miss data for the three independent inspectors used in the study were combined ('pooled') for the purposes of the analysis to provide a more robust data set. Considerable care was taken to ensure that the original data had been correctly transcribed when creating the input data files. The analysis software implements currently accepted statistical techniques to the analysis of hit/miss data compared to the overlapping sampling method used in the original work. The false-call rate was not recorded in the original report [15] but subsequently indicated as less than 5% [30]. The DSTO reanalysis did not take into account the false-call rate.

In the original study [15], fatigue cracks were grown from EDM starter notches in three-point bending. When cycling had been completed, the surface material containing the EDM starter notches was machined away to leave a flat specimen containing fatigue cracks in known locations. In the 'as-machined' state, there is a possibility that the cold-worked surface layer may interfere with crack detectability. Hence, the specimens were subsequently etched and subjected to a tensile proof load of 80% of the material yield strength. The authors indicated that such an etching and proof load treatment simulated state-of-the-art industry practices of the time. Inspections were made for both the as-machined specimens and after etching and proof loads.

* For Sequence 1 (as machined) 53 out of 146 crack lengths were incorrect. For Sequence 3 (after etch and proof load), 53 out of 176 crack lengths were incorrect.

One complication encountered in the original 1976 work was that a number of 'unintentional' cracks were observed in addition to those 'planned' cracks which had been initiated from the EDM slots. It appears that, while the locations, dimensions and hit/miss data for these unintentional cracks were recorded, the authors did not include them in their analysis 'due to problems incurred in matching, verifying and correlating actual data to NDT observations' [15]. For the sake of completeness, the reanalysis is carried out for (i) the planned set of cracks (i.e., the 111 cracks* with crack identification numbers < 200), as well as (ii) all cracks (planned and unintentional) totalling 142 cracks in the as-machined specimens and 176 cracks in the etched and proof-loaded specimens.

The results of the analysis are given in Table C1 for inspections carried out with the specimens in the 'as machined' state and after subsequent 'etching and proof load'. The POD and lower 95% confidence curves are plotted in Figures C1 – C4. As discussed above, results are given for the planned crack population and the population including unintended cracks.

For the record, the results of the DSTO reanalysis of the Rummel et al data [15] are compared with those presented in the *NDE Data Capabilities Handbook* [30]. As mentioned previously, the Handbook analysis is flawed because it was based on incorrect defect sizes. There are also two important differences between the results presented in the Handbook and those in the original study by Rummel et al. [15].

(i.) Rather than presenting pooled data, the *NDE Capabilities Data Handbook* reports the POD analyses for each individual inspector.

(ii.) The analysis in the Handbook uses data for all cracks (i.e., unintentional and planned cracks) rather than the subset of planned cracks in the original work.

The comparison between the DSTO reanalysis and the Handbook results is presented in Table C2 and Table C3.

Table C1 Results of the reanalysis of the Rummel et al. 1976 study of the reliability of fluorescent MPT for crack detection in 4340 steel [15]. These data are also plotted in Figures C1–C4.

	a_{90} (mm)	$a_{90/95}$ (mm)
Planned cracks		
As machined	4.02	5.14
After etching and proof load	2.01	2.48
All cracks		
As machined	4.93	7.26
After etching and proof load	4.57	6.21

* This may only be a subset of the planned number of cracks as Rummel et al [12] (page 9) state that 128 planned flaws were introduced into the specimens.

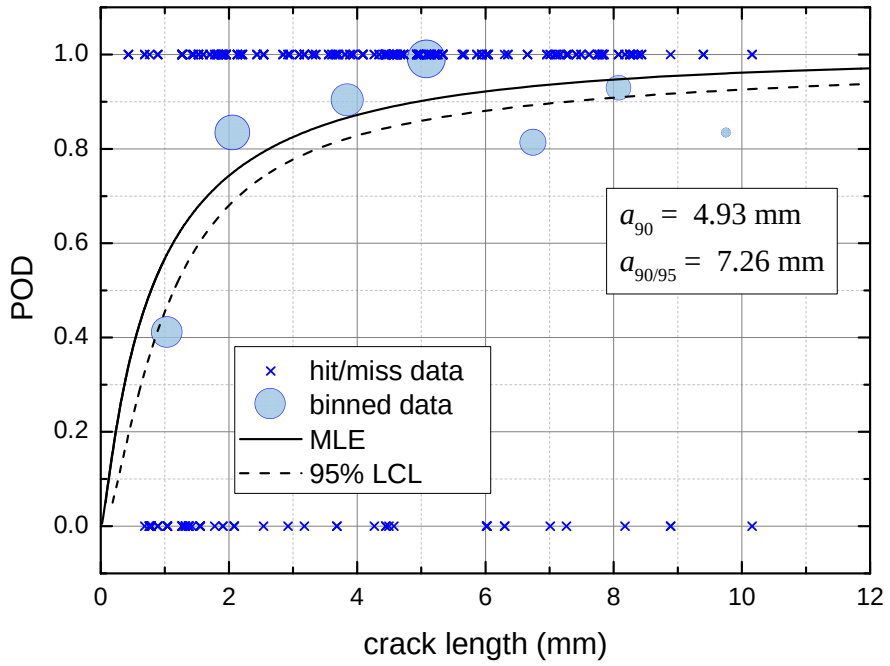

Figure C1 Reanalysis of Rummel et al. (1976) data for fluorescent MPT of fatigue cracks in as machined 4340 steel plates [15]. The hit/miss results for all 142 cracks are included. The POD curves were generated after pooling the results for all three inspectors.

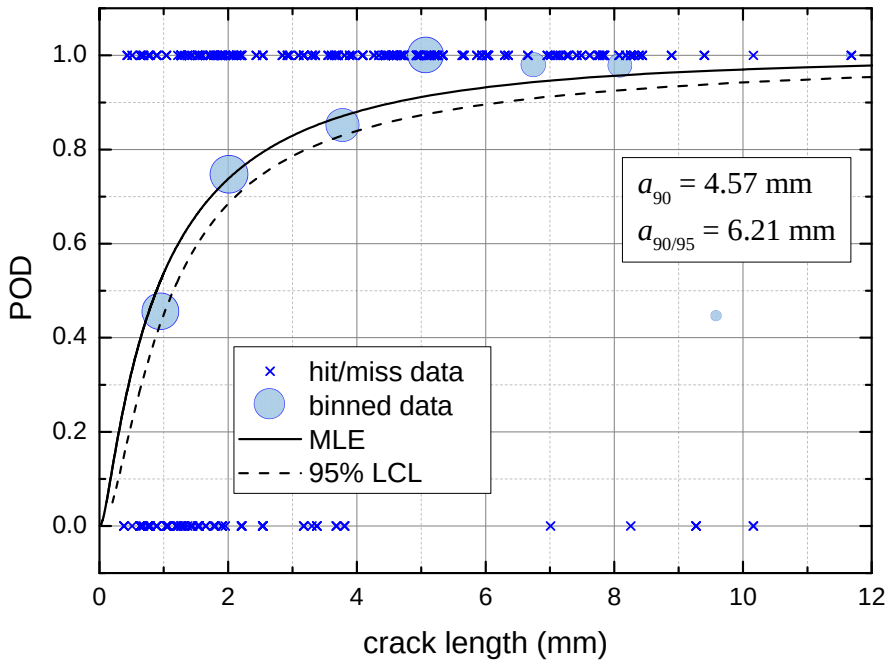

Figure C2 Reanalysis of Rummel et al. (1976) data for fluorescent MPT of fatigue cracks in 4340 steel plates after etching and proof load [15]. The hit/miss results for all 176 cracks are included. The POD curves were generated by pooling the results for all three inspectors.

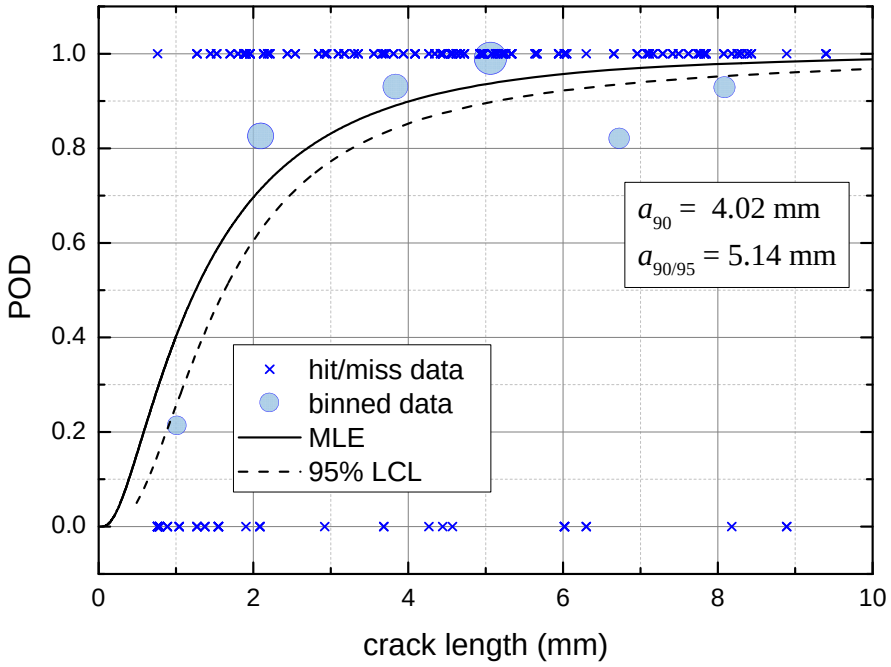

Figure C3 Reanalysis of Rummel et al. (1976) data for fluorescent MPT of fatigue cracks in as machined 4340 steel plates [15]. The hit/miss results included only the 111 intentional cracks. The POD curves were generated by pooling the results for all three inspectors.

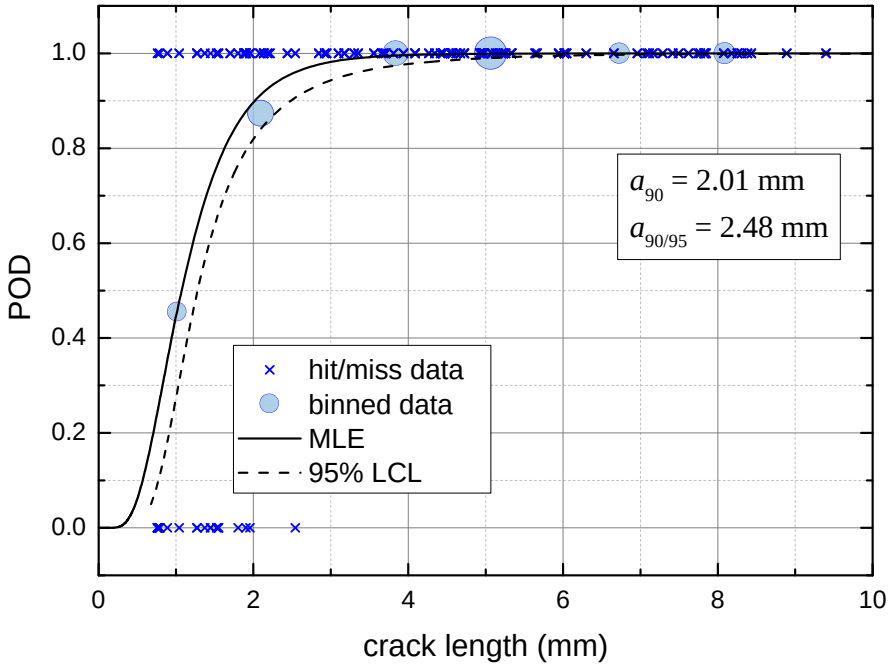

Figure C4 Reanalysis of Rummel et al. (1976) data for fluorescent MPT of fatigue cracks in 4340 steel plates after etching and proof load [15]. The hit/miss results included only the 111 intentional cracks. The POD curves were generated by pooling the results for all three inspectors.

Table C2 Reliability of fluorescent MPT for crack detection in 4340 steel [15]. Comparison between the present DSTO reanalysis and the NDE Data Capabilities Handbook for as-machined specimens.

	a_{90} (mm)		Fraction of hits	
	Handbook	DSTO	Handbook	DSTO
All cracks				
Inspector 1 (A)	17.56	7.43	108/142	108/142
Inspector 2 (B)	10.10	4.14	120/142	120/142
Inspector 3 (C)	9.31	3.54	120/142	123/142
Planned cracks				
Inspector 1 (A)	–	3.61	–	94/111
Inspector 2 (B)	–	4.51	–	92/111
Inspector 3 (C)	–	3.95	–	95/111

Table C3 Reliability of fluorescent MPT for crack detection in 4340 steel [15]. Comparison between the present DSTO reanalysis and the NDE Data Capabilities Handbook for specimens after etching and proof load.

	a_{90} (mm)		Number of hits	
	Handbook	DSTO	Handbook	DSTO
All cracks				
Inspector 1 (A)	6.59	5.43	124/176	124/176
Inspector 2 (B)	2.68	4.90	142/176	141/176
Inspector 3 (C)	3.04	2.87	154/176	154/176
Planned cracks				
Inspector 1 (A)	–	2.40	–	94/111
Inspector 2 (B)	–	1.53	–	106/111
Inspector 3 (C)	–	1.36	–	107/111

Appendix D: Reanalysis of NRC IAR studies

D.1 NATO AGARD Round Robin Study (1994)

As part of a large AGARD collaborative program led by NRC IAR Canada, a series of POD trials was carried out to examine the reliability of inspection for aircraft gas turbine engine components [23-25]. Six laboratories in four NATO countries participated in the trials, three laboratories (denoted as Organisation I, II and III) contributed to the MPT element of the program. The organisations were not identified.

The test components consisted of discs and spacers from retired J95-CAN-40 engines which contained in-service low-cycle fatigue cracks at a series of boltholes. The disc and engine material was AM355, a precipitation hardened martensitic stainless steel. The inspections were based on MIL-STD-1949A [51], now superseded but current at the time. Organisation III performed the inspections using wet fluorescent particle MPT with two magnetisation shots at 90° apart. Detailed information on the inspection procedures was not provided by the two other organisations. Only the bolthole region was inspected. Following the inspections the disc and spacer boltholes were subjected to destructive testing to determine the crack size and type.

The reported results are summarised in Table D1. In this round-robin exercise, a common set of test specimens was inspected with the exception of the engines spacers, for which Organisation I and III inspected a different subset of the spacers. This leads to the different numbers of cracks and different crack population inspected by each organisation. In the case of Organisation III, the subset of spacers inspected contained only small cracks ($a < 1.2$ mm and generally less than 0.5 mm) and no hits were recorded. In this study, all cracks larger than 2.3 mm were detected with the exception of one crack 4.6 mm long which was missed by Organisation I. Organisation III detected a larger number of small cracks but with a high false-call rate. Organisation I, on the other hand, had a zero false-call rate but missed a number of small cracks as well as a 4.6 mm crack. The reported a_{90} values for the three organisations vary from 1.8 mm to 3.3 mm.

Table D1 Reported outcomes from the NATO AGARD POD trials: MPT [23,24].

	Number of cracks	a_{90} (mm)	False-call rate†
Organisation I	285	3.3	0%
Organisation II	404	1.8	4.7%
Organisation III	207	2.6*	10.4%

* Derived from probability of indication curve.
† Rate is the number of known false calls/number of uncracked holes.

The method used to calculate the lower 95% confidence level (and hence $a_{90/95}$) in the NATO study has subsequently been shown to be deficient [2,34]. The POD data were therefore reanalysed by DSTO. The raw hit/miss data were taken from the *NDE Capabilities Data Handbook* CD [30] and cross-checked against the original reports [23,24]. The Handbook did not contain a listing for the spacer inspections by Organisation III (which contained no hits) so these data were extracted manually from the original report. No discrepancies were found between the Handbook listing of the raw data and those in the original report (Table II and Appendix C of [24]).

In the DSTO reanalysis, the disc and spacer inspection results were pooled for each organisation, as in the original analysis by NRC IAR. This differs from the reanalysis presented in the *NDE Capabilities Data Handbook*, for which the disc and spacer data for each organisation were treated separately. Pooling the data for discs and spacers, which are both essentially bolthole inspections, increases the number of inspections and results in a more robust data set. The results of the DSTO reanalysis are given in Table D2, where three different methods for pre-processing the data have been used. The first column gives the calculated a_{90} and $a_{90/95}$ values without any pre-processing, i.e., all the hit/miss data are included in the analysis. The a_{90} values agree well with those in the original NRC IAR report (reproduced in Table D1) which also did not include any pre-processing. The $a_{90/95}$ values from the DSTO reanalysis are smaller than those originally reported (Table II of [24]) for which the analysis methods used to calculate confidence levels were deficient.

In later work (Section D.2), Forsyth and Fahr [26,27] consider the effect of rogue data on the estimation of POD for similar discs and spacers. The specimens used in the NATO AGARD round robin and the later Canadian study contained a large number of very small cracks ($a < 0.3$ mm). These cracks were so small that MPT was not capable of detecting them consistently. The hit indications were rare, and considered most likely to be false indications fortuitously occurring at the same locations as very small cracks. When fitting a two-parameter log-normal POD curve to hit/miss data, any false indications at very small crack lengths will have a disproportionate effect on the resulting POD curve. Because of the limitations of the two-parameter POD model, a false indication for a very small crack can influence the fitted POD curve as severely as missing a large crack. To overcome this limitation without resorting to fitting a four-parameter POD curve, Forsyth and Fahr pre-process the hit/miss data in [26,27] to exclude all hits for crack lengths below 0.3 mm to allow analysis using the conventional two-parameter POD curve. Such an approach tends to result in a more steeply rising POD curve and often reduces the estimated a_{90} value.

The effect of pre-processing the NATO AGARD hit/miss data to exclude hits for small cracks ($a < 0.3$ mm) is shown in Table D2. In all cases, excluding hits for crack lengths less than 0.3 mm leads to a decrease in the calculated a_{90} values. The most striking example of the effect of such pre-processing is for Organisation I: excluding just the one data point (a 'rogue hit' for a crack 0.09 mm long) leads to a decrease in a_{90} from 3.30 mm to 2.39 mm.

Rather than selectively excluding just the hits for crack lengths less than 0.3 mm, another option is to exclude all data (hits and misses) for such very small crack lengths. In this way there is no bias. The results of pre-processing the NATO AGARD data to exclude all small cracks (a < 0.3 mm) are also shown in Table D2. The differences in the a_{90} and $a_{90/95}$ values calculated using the two different options for excluding data for small crack lengths are negligible. In the DSTO reanalysis of the NATO AGARD study, the preferred method for pre-processing is to exclude all data for crack lengths below 0.3 mm. The corresponding POD curves are shown in Figures D1 – D3.

The unique feature of the NATO AGARD round robin was the use of retired engine components for the POD study. The reliability of inspection for in-service fatigue cracks was measured rather than for laboratory-grown fatigue cracks. The trial is therefore more representative of in-service inspections. The drawback in using retired components is that the crack population cannot be controlled to the same degree as for laboratory-grown cracks. This is illustrated by the spacer specimens inspected by Organisation III where the subset of spacers supplied contained only small cracks, none of which were detected.

A second feature of the NATO AGARD trials was that three distinct crack geometries were grouped together in the study: mid-bore cracks, corner cracks and through cracks. As a consequence, the hit/miss data for smaller crack lengths (a < 1.5 mm) are dominated by mid-bore and corner cracks whereas the large crack data (a > 1.5 mm) are dominated by through-thickness cracks.

Table D2 DSTO reanalysis of NATO AGARD POD trials using different pre-processing of data

	All data		Exclude hits < 0.3mm			Exclude data < 0.3mm		
	a_{90} (mm)	$a_{90/95}$ (mm)	a_{90} (mm)	$a_{90/95}$ (mm)	# points excluded	a_{90} (mm)	$a_{90/95}$ (mm)	# points excluded
Org. I	3.30	4.66	2.39	2.96	1	2.39	2.96	75
Org. II	1.76	2.28	1.44	1.77	4	1.46	1.81	152
Org. III	2.58	4.61	1.53	2.20	6	1.54	2.25	96

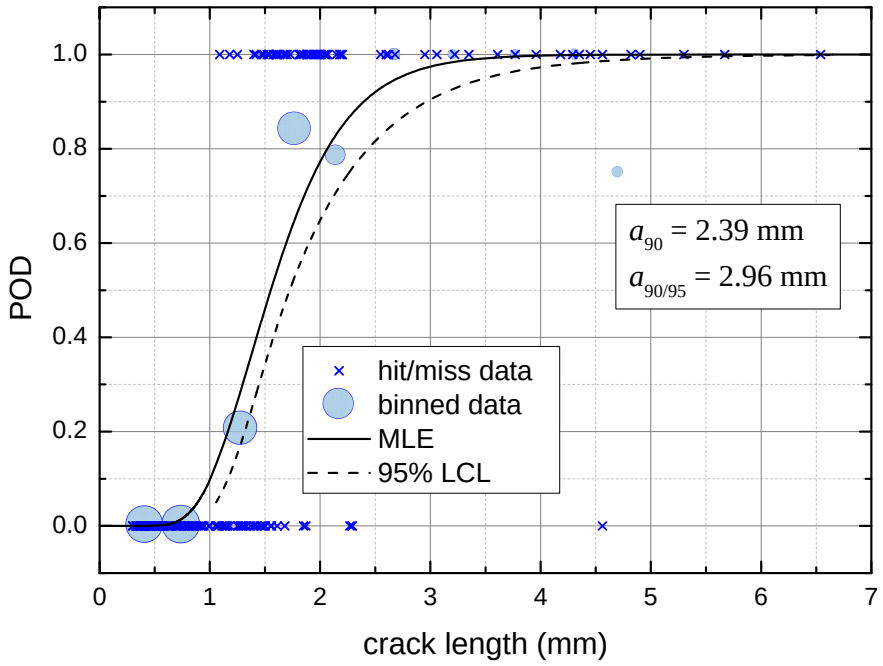

Figure D1 Reanalysis of Fahr et al. (1994) data (organisation I) for MPT of fatigue cracks at holes in engine discs and spacers made from AM 355. Crack lengths less than 0.3 mm were excluded.

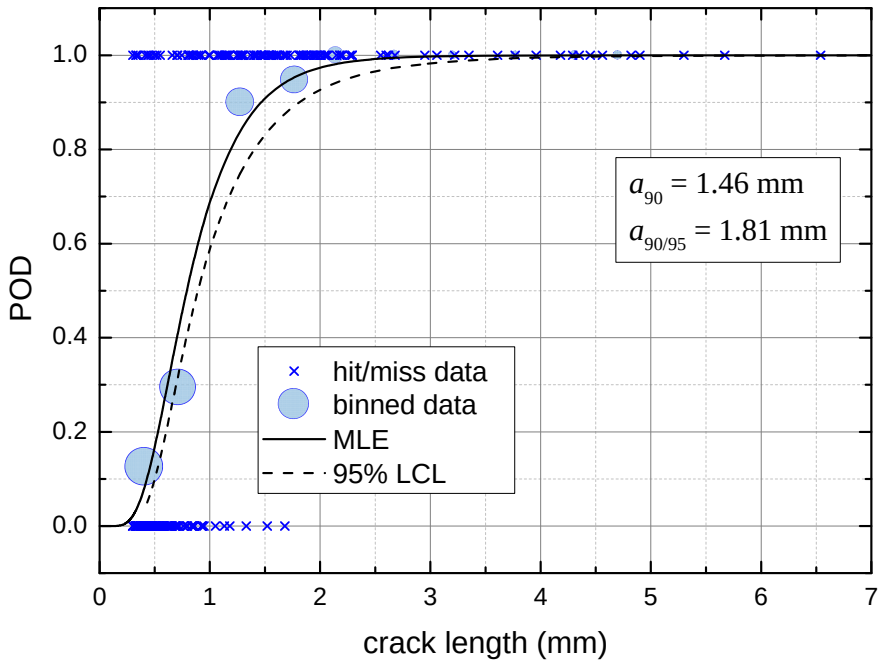

Figure D2 Reanalysis of Fahr et al. (1994) data (organisation II) for MPT of fatigue cracks at holes in engine discs and spacers made from AM 355. Crack lengths less than 0.3 mm were excluded.

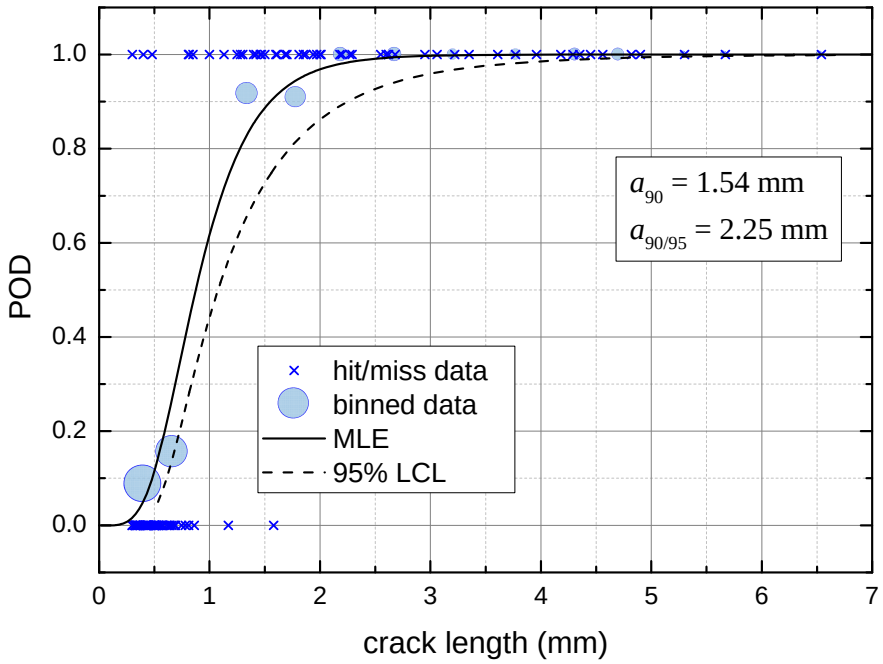

Figure D3 Reanalysis of Fahr et al. (1994) data (organisation III) for MPT of fatigue cracks at holes in engine discs and spacers made from AM 355. Crack lengths less than 0.3 mm were excluded.

D.2 Canadian Study (1996)

NRC IAR coordinated a further round-robin POD study of retired J95-CAN-40 engine discs following the NATO AGARD trials [26]. Six Canadian organisations participated in this program. MPT was conducted by only one organisation. Boltholes in ten retired engine discs were inspected and crack dimensions determined by destructive testing. The false-call rate (13%) was larger than the highest false-call rate in the NATO-AGARD round robin. The largest crack which was missed was 2.04 mm in length.

In the original analysis, Forsyth and Fahr [26] pre-process the data to exclude any hits for very small cracks ($a < 0.3$ mm). The authors argue that hits at very small crack lengths are most likely false indications and, as discussed in Section D.1, including such rogue data has a disproportionate effect when fitting a two-parameter cumulative log-normal POD curve. As the original method used to calculate the lower 95% confidence level has subsequently been shown to be deficient [2,34], the POD data were again reanalysed by DSTO. The raw hit/miss data were extracted from the original report (Appendix C of [26]) and were cross-checked against the listing provided in the *NDE Capabilities Data Handbook* [30]. While no discrepancies were found in the listings of the raw data in the Handbook, the analysis method in the Handbook included all data without any pre-processing. Furthermore, the Handbook summary reports a calculated value of $a_{90} = 4.33$ mm compared with $a_{90} = 4.7$ mm attributed to Forsyth and Fahr [26]. However, the latter value could not be found in the original NRC IAR report.

Table D3 DSTO reanalysis of NRC IAR POD trial using different pre-processing of data (320 cracks)

All data		Exclude hits < 0.3 mm			Exclude data < 0.3 mm		
a_{90} (mm)	$a_{90/95}$ (mm)	a_{90} (mm)	$a_{90/95}$ (mm)	# points excluded	a_{90} (mm)	$a_{90/95}$ (mm)	# points excluded
5.31	10.34	2.43	3.34	9	2.44	3.37	109

The results of the DSTO reanalysis are presented in Table D3, where three different pre-processing methods have been used. Pre-processing the data to exclude all hits for cracks less than 0.3 mm leads to a significant reduction in a_{90} and $a_{90/95}$, as observed in the reanalysis of the NATO AGARD round-robin data (Table D2). There is a negligible difference in the a_{90} and $a_{90/95}$ values calculated by excluding all data for $a < 0.3$ mm rather than just hits for $a < 0.3$ mm. The a_{90} value (2.44 mm) is comparable with the values calculated by DSTO for the NATO AGARD round robin trials (Table D2). The corresponding POD and lower 95% confidence curves are shown in Figure D4.

The a_{90} value (2.44 mm) calculated via the DSTO reanalysis, excluding hits for $a < 0.3$ mm, is in good agreement with the value $a_{90} = 2.4$ mm in the original report [26]*. The $a_{90/95}$ value obtained by the DSTO analysis is larger than the value $a_{90/95} = 2.67$ mm quoted in the original report, for which the original analysis is deficient.

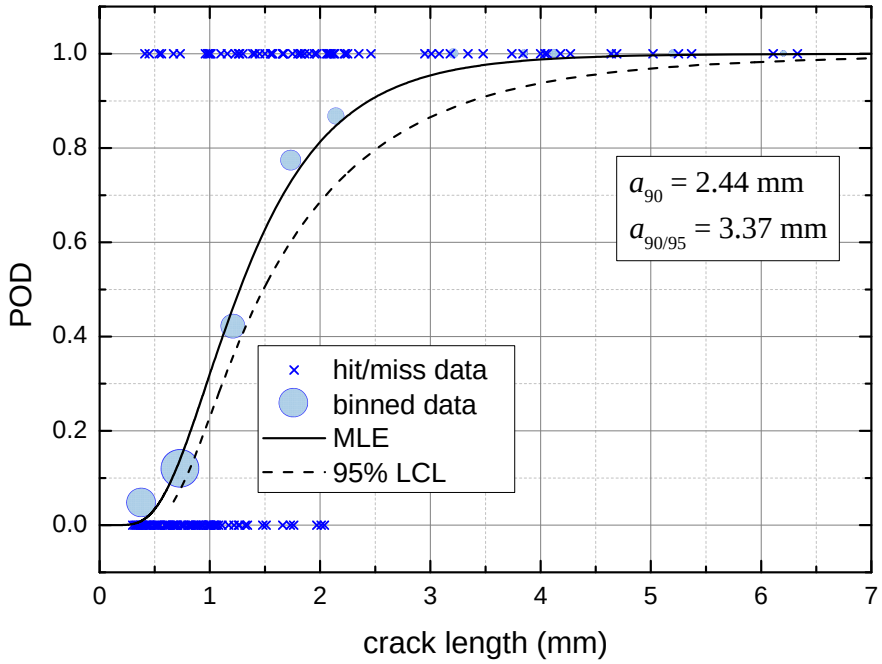

Figure D4 Reanalysis of Forsyth et al. (1996) data (Org. D) for MPT of fatigue cracks at holes in engine discs made from AM 355. Crack lengths less than 0.3 mm were excluded.

* The a_{90} value is obtained from the probability of indication curve (Forsyth and Fahr [27], Figure 12) rather than the POD curve because the DSTO reanalysis did not include a false call analysis.

As was the case in the NATO AGARD trials (Section D.1), Forsyth and Fahr [26] group together mid-bore cracks, corner cracks and through-thickness cracks for the purposes of the analysis. Thus the POD curve is dominated by corner- and mid-bore cracks for smaller crack lengths (a < 1.5 mm) and by through-thickness cracks for larger crack lengths (a > 1.5 mm)*.

* For through-thickness cracks and corner cracks, a is defined in the original report as the length of the exposed crack on either the top or bottom surface of the disc or spacer. For mid-bore cracks, a is taken as the full crack length within the bore of the hole.

Appendix E: Statistical inferences on a_{90}

In the meta-analysis presented in Section 4.6, a number of statistical measures relating to the reliability of MPT are discussed. These statistical inferences assume that the a_{90} values derived from a reanalysis of the MPT reliability literature follow a log-normal distribution. For completeness, the results and closed-form expressions used to calculate the relevant statistics are reproduced below.

E.1 Maximum likelihood estimates

The log-normal probability density function is defined as follows:

$$P(x;M,S)=\frac{1}{\sqrt{2\pi}\,x\,S}\exp\left[-\frac{\left(-M+\ln x\right)^2}{2S^2}\right],\quad x>0, \tag{E1}$$

where the parameters M and S^2 are the mean and variance of the normal distribution from which the log-normal distribution is derived. In the present case, the variable x corresponds to a_{90}.

Given a set of x values (x_1, x_2, … , x_n) which follow a log-normal distribution, the maximum likelihood estimates for M and S^2 are given by the expressions

$$\hat{M}=\frac{1}{n}\sum_{i=1}^{n}\ln\left(x_i\right), \tag{E2}$$

$$\hat{S}^2=\frac{1}{n}\sum_{i=1}^{n}\left[\ln\left(x_i\right)-\hat{M}\right]^2. \tag{E3}$$

Equations (E2) and (E3) are derived by maximising the log-likelihood function*

$$\ln\mathcal{L}=\sum_{i=1}^{n}\ln P\left(x;M,S\right) \tag{E4}$$

with respect to M and S^2.

Using the known properties of the log-normal distribution, the MLE values for the median, mean and mode values of a_{90} are given by the expressions

$$\hat{x}_{median}=\exp\left(\hat{M}\right),\quad \hat{x}_{mean}=\exp\left(\hat{M}+\hat{S}^2/2\right),\quad \hat{x}_{mode}=\exp\left(\hat{M}-\hat{S}^2\right), \tag{E5}$$

and the variance in a_{90} is

* The MPT data set does not contain any right censored data so this simpler expression applies, compared to the log-likelihood function used for similar analysis of LPT data [2].

$$\hat{\sigma}^2 = \exp\left(2\hat{M} + \hat{S}^2\right)\left[\exp\left(\hat{S}^2\right) - 1\right]. \quad \text{(E6)}$$

The predicted a_{90} percentiles can also be derived from the properties of the log-normal distribution so that the MLE P^{th} percentile of the a_{90} distribution is given by

$$\hat{x}_{P\%} = \exp\left[\hat{M} + z_p \hat{S}\right], \quad \text{(E7)}$$

where $z_p = \sqrt{2}\,\mathrm{erf}^{-1}(-1 + 2p)$, $p = P/100$ and erf^{-1} denotes the inverse error function. In the case of the 90th percentile, $p = 0.9$ and $z_p = 1.28155$.

Substituting the six selected values (2.32 mm, 2.01 mm, 2.39 mm, 1.46 mm, 1.54 mm and 2.44 mm) for a_{90} from Section 4.6 into (E2) and (E3) gives the MLE parameters for the assumed a_{90} log-normal distribution

$$\hat{M} = 0.685535, \qquad \hat{S}^2 = 0.0433896. \quad \text{(E8)}$$

The resulting distribution is plotted in Figure E1 and is compared with the corresponding distribution of a_{90} values for LPT derived by Harding and Hugo [2] using the eight trial results from the NRC IAR LPT studies. The variability in the MPT a_{90} values is significantly smaller than that reported for the LPT POD studies [2].

Having obtained the MLE estimates for M and S^2, the MLE estimates for the mean, median and variance of the MPT a_{90} values can be calculated by substituting (E8) into (E5) and (E6). The results are given in Table E1 and are compared with the sample statistics for the data set. The mean and variance obtained using the MLE method with an assumed log-normal distribution agree well with the sample statistics. The difference in the median values arises from the coarseness inherent in calculating the sample median, which relies on averaging the middle values in the sorted list without regard to the shape of the underlying distribution.

The predicted 90th percentile values for a_{90} can be calculated by substituting (E8) into (E7) with $p = 0.9$ to obtain

$$\hat{x}_{90\%} = 2.59\,\mathrm{mm}. \quad \text{(E9)}$$

On this basis, 90% of MPT implementations are expected to achieve $a_{90} < 2.6$ mm.

Table E1 Descriptive statistics for MPT a_{90} values. Results obtained using simple sample statistics are compared with those calculated using the MLE method assuming a log-normal distribution (all dimensions in mm).

	Median	Mean	Variance	Maximum	Minimum
Simple statistics	2.17	2.03	0.189	$a_{90}^{max} = 2.44$	$a_{90}^{min} = 1.46$
Log-normal MLE	1.98	2.03	0.182		

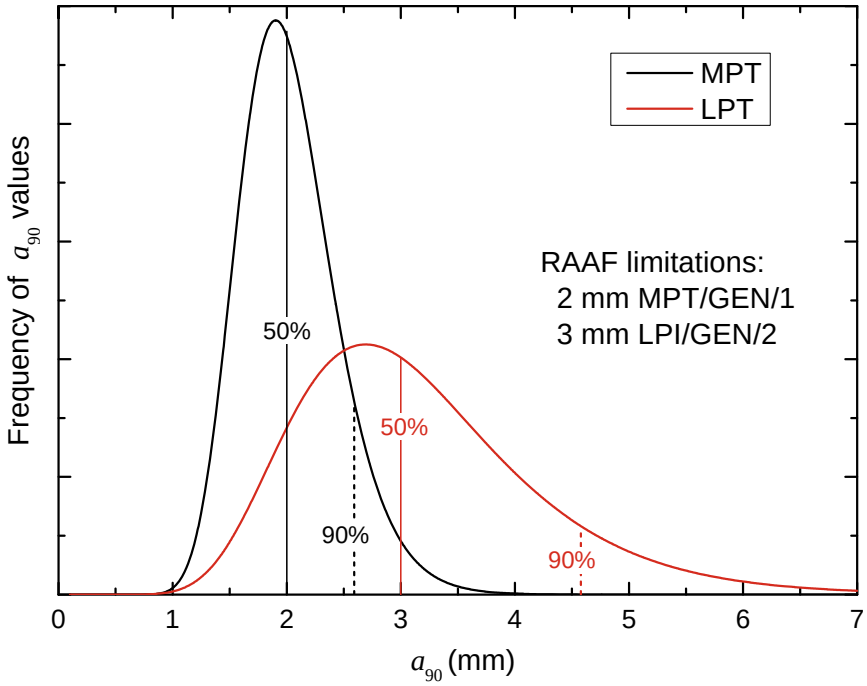

Figure E1 Assumed log-normal distribution of a_{90} values from POD trials obtained using the MLE method. The results for MPT are compared with those obtained for LPT by Harding and Hugo [2]. In both cases, the RAAF standard limitation corresponds to the median of the distribution. The LPT data show considerably more variability than the MPT data.

E.2 Confidence limits on the median a_{90}

The median a_{90} = 2.0 mm calculated in the previous section is the best statistical estimate obtained from the data set. To account for the randomness arising from the finite size of the sample, confidence limits can be placed on the range in which the true median lies. In the analysis presented below, we consider the two-sided confidence limit on the median and seek the range $x_{CL(-)} < x_{median} < x_{CL(+)}$ which brackets the median with 90% statistical confidence.

The confidence limits are calculated by seeking the maximum and minimum values of the median $x_{median} = e^M$ subject to the constraint

$$Q_2 - \gamma = 0 \tag{E10}$$

where Q_2 is related to the likelihood ratio and is defined by

$$Q_2 = -2\ln\left(\mathcal{L}/\hat{\mathcal{L}}\right) = -2\sum_{i=1}^{n}\left[\ln P\left(x;M,S\right) - \ln P\left(x;\hat{M},\hat{S}\right)\right] \tag{E11}$$

and $\gamma = 2\,[\text{erf}^{-1}(1-\alpha)]^2$ is related to the required confidence limits through the parameter α. For example, for the two-sided 90% confidence limits, $\alpha = 0.1$ and so $\gamma = 2.705$.

A closed-form expression for Q_2 can be derived for a log-normal distribution by substituting (E1) and (E4) into (E11) and simplifying using the relations (E2) and (E3) for the MLE estimates of M and S, so that

$$Q_2 = 2n\left[\frac{(M-\hat{M})^2+\hat{S}^2-S^2}{2S^2}+\ln\left(S/\hat{S}\right)\right]. \tag{E12}$$

Substituting (E12) into (E10), it can be shown, after some manipulation, that the constraint (E10) requires the values of M and S to lie on the oval-shaped curve

$$v^2 = u^2 - 1 + 2u^2\left(\frac{\gamma}{2n}-\ln u\right) \quad \text{where} \quad v=\frac{(M-\hat{M})}{\hat{S}}, \quad u=\frac{S}{\hat{S}}. \tag{E13}$$

The maximum and minimum values of the median $x_{median} = e^M$ subject to the constraint (E13) can then be found by seeking the maximum and minimum values of M on the curve. These extrema can be determined either by inspection or by differentiating (E13) with respect to x, setting the derivative to zero and solving the resulting equation to give

$$M_{\pm} = \hat{M} \pm \hat{S}\sqrt{\exp\left(\frac{\gamma}{n}\right)-1}, \quad S_{\pm} = \hat{S}\exp\left(\frac{\gamma}{2n}\right), \tag{E14}$$

where M_+ and M_- are the respective maximum and minimum values of M.

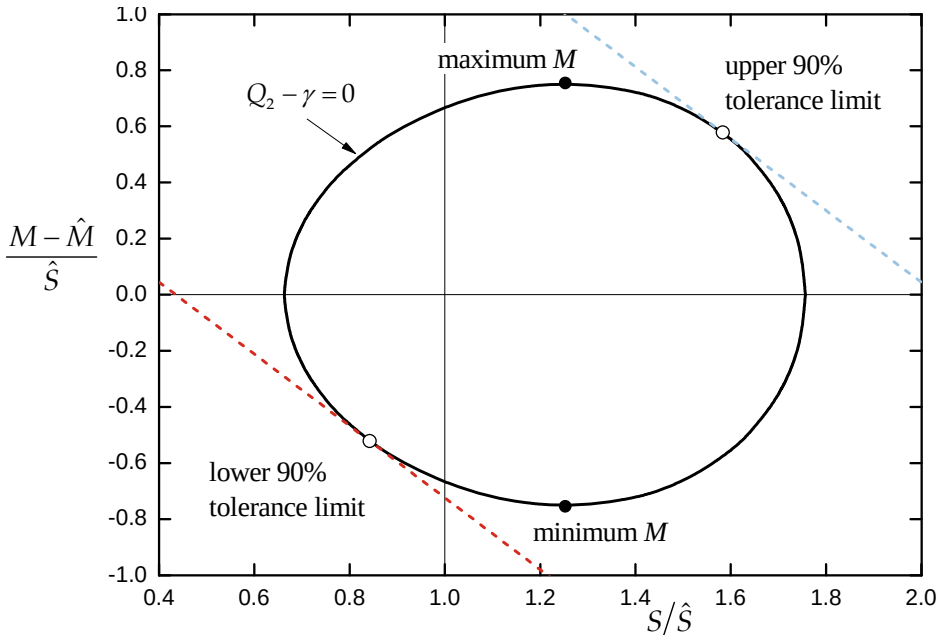

Figure E2 Locus of the constraint $Q_2-\gamma = 0$ for $\gamma = 2.705$ and $n = 6$ from equation E13. The two-sided 90% confidence limits on the median a_{90} are deduced from the maximum and minimum values of M (closed circles). The dashed lines indicate the construction used to obtain the two-sided confidence limits on the 90th percentile a_{90}.

The corresponding (two-sided) upper and lower confidence limits on the median are then

$$x_{CL(\pm)} = \exp\left[\hat{M} \pm \hat{S}\sqrt{\exp\left(\frac{\gamma}{n}\right)-1}\right]. \tag{E15}$$

Substituting the MLE estimates (E8) for M and S into (E15), with n = 6 and γ = 2.705, the two-sided 90% confidence limits on the median for the MPT data set are $x_{CL(-)}$ = 1.696 mm and $x_{CL(+)}$ = 2.323 mm respectively.

E.3 Confidence limits on the 90th percentile value of a_{90}

Similar to the median a_{90}, statistical confidence (or tolerance) limits on the predicted 90th percentile value for a_{90} can also be calculated by employing the likelihood ratio Q_2. Thus, we seek the maximum and minimum values for the 90th percentile value of a_{90}

$$x_{90\%} = \exp\left(M + z_{0.90}\, S\right), \tag{E16}$$

subject to the constraint $Q_2 - \gamma = 0$. The maximum and minimum values for $x_{90\%}$ will occur when the argument of the exponential

$$c = M + z_{0.90}\, S, \tag{E17}$$

is either a maximum or minimum.

In graphical form, the solution to this constrained optimisation problem can be found by seeking the values of c for which the line

$$v = -z_{0.90}\, u + (c - \hat{M})/\hat{S} \tag{E18}$$

touches the oval defining the locus of normalised (M, S) values satisfying (E13). Unlike the confidence limits on the median (Section E2), it does not appear possible to derive a closed-form expression for the tolerance limits and the solution must be found using numerical methods.

The solution is depicted in Figure E2, where the two points of intersection are found to be (u = 0.8421, v = – 0.5221) and (u = 1.5829, v = 0.5777) for the particular case where 90% statistical confidence is sought (γ = 2.705) and n = 6 corresponding to the MPT data set.

Substituting the (u, v) values for the points of intersection into (E18) together with $\hat{M}$ and $\hat{S}$ from (E8) for the MPT data set, gives the two required extreme values for c, c = 0.8016 and c = 1.2284. Hence, substituting these extreme values for the argument of the exponential (E17) into (E16), gives the 90% confidence limits on $x_{90\%}$ of 2.229 mm and 3.416 mm respectively.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Review of Literature on Probability of Detection for Magnetic Particle 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)

S.K. Burke and R.J. Ditchburn

5. CORPORATE AUTHOR

DSTO 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-2794	AR-015-495	Technical Report	January 2013

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2012/1004787/1	AIR 07/101	DGTA ADF	50	51

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

Nondestructive testing, Structural integrity, Reliability, Probability

19. ABSTRACT

A critical review of the available literature concerning the minimum reliably detectable defect size a_{NDI} for magnetic particle testing (MPT) in aerospace applications is presented. Four probability of detection (POD) studies relevant to detection of fatigue cracks in aircraft components were found over the period 1968 to 2011. As the statistical methods used in these four previous studies were either outdated or otherwise deficient, the original data were reanalysed using currently accepted techniques. A meta-analysis of the results is presented, with emphasis on statistical inferences for the defect size expected to be detected with 90% POD. It is shown that the minimum reliably detectable defect size a_{NDI} = 2.0 mm currently specified by the Royal Australian Air Force for wet fluorescent magnetic particle inspection using the continuous method is consistent with estimates of the average performance of MPT derived from the reanalysis of the literature.