NASA NDE FRACTURE CRITICAL DETECTABLE FLAW SIZES HISTORY AND METHODOLOGY

Peter A. Parker[1], Ajay Koshti[2], David S. Forsyth[3], Michael W. Suits[4], James L. Walker[5] and William H. Prosser[6]

[1] NASA Langley Research Center, Hampton, Virginia, peter.a.parker@nasa.gov
[2] NASA Johnson Spaceflight Center, Houston, Texas
[3] NDT Analysis, St. John, U.S. Virgin Islands
[4] NASA Marshall Spaceflight Center, Huntsville, Alabama
[5] NASA Marshall Spaceflight Center, Huntsville, Alabama
[6] NASA Engineering and Safety Center, Hampton, Virginia

Abstract: NASA requires that NDE methods and inspectors demonstrate 90% Probability of Detection (POD) with 95% confidence for critical flaw sizes when inspecting fracture critical metallic components. NASA addresses the known variability of NDE inspector capability in two ways. The first, Special NDE, requires that every inspector demonstrate the required 90/95 POD, which is resource intensive. The second approach is Standard NDE for which conservative flaw sizes for different NDE methods are provided such that it is expected that most properly trained inspectors will exceed the 90/95 POD requirement. As such, individual POD demonstration testing is not required.

The origin of NASA Standard NDE dates to the start of the Space Shuttle Program in the early 1970's. In the first study to quantitatively assess NDE methods and inspectors, the performance of multiple inspectors was evaluated for different NDE methods using a large set of fatigue cracked specimens. A rudimentary POD analysis was performed to estimate the 90/95 POD flaw size for each inspector for each method. Additionally, the average and standard deviation of the 90/95 POD flaw size across the multiple inspectors was calculated to estimate the flaw size for which 95 percent of inspectors would provide the 90/95 POD capability. These estimated 90/95/95 POD flaw sizes evolved into the NASA Standard NDE flaw sizes still in use for structural analysis five decades later.

The methodology for performing Standard NDE POD studies was never documented in NASA requirements. Furthermore, POD analysis methods have significantly evolved since this seminal study. Likewise, NDE methods have improved and there has been a push to reassess Standard NDE flaw sizes for existing methods, and to develop Standard NDE flaw sizes for new methods such as digital radiography. In this study, a Standard NDE POD methodology was developed and baselined using the historical data. This reanalysis of the historical data identified several deficiencies in the original test plan as well as an overall lack of conservatism in the estimated 90/95/95 POD flaw sizes. The results of this historical review and the new methodology are being incorporated into an update of NASA NDE POD requirements.

Keywords: Standard NDE, Inspector Variability, POD reliability <<need more>>

INTRODUCTION

A NASA Standard nondestructive evaluation (NDE) flaw size is intended to represent the largest flaw size that may be missed by most qualified inspectors using a specific NDE method. Therefore, the Standard NDE flaw size is assumed to exist in the worst-case location and orientation on a part in the fracture analysis assessment of component lifetime to show conformance to "NASA-STD-5019A Fracture Control Requirements for Spaceflight Hardware" [1] requirements. The tabulated Standard NDE flaw sizes in NASA's "NASA-STD-5009B Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components" [2] are used in the majority of NASA's human spaceflight system designs. The primary benefit of tabulating Standard NDE flaw sizes for commonly used inspection methods is that it avoids the requirement for individual inspector probability of detection (POD) demonstrations, required for NASA Special NDE, which can be resource intensive.

In 2022, NASA embarked on an expedition to rigorously trace the Standard NDE flaw sizes in NASA-STD-5009B, a challenging 50-year retrospective survey, which is documented in "A Survey of NASA Standard Nondestructive Evaluation (NDE)" [3]. There were significant findings regarding the POD study's flaw fabrication, flaw size distribution, and the rudimentary POD analysis methods used in the 1970s that resulted in non-conservative flaw sizes compared to modern analysis using MIL-HDBK-1823A [4] methodology.

While the legacy Standard NDE flaw sizes have been successfully applied to many NASA programs and projects since the Space Shuttle Program (SSP), in most cases they do not represent what they were reported to be, namely the flaw size that provides 90% POD with 95% confidence for 95% of inspectors, known as the a90/95/95 flaw size. Furthermore, the methodology for performing and analyzing a Standard NDE POD study was not codified in any of the NASA requirements documents. Thus, there was no established methodology to develop Standard NDE flaw sizes for new NDE methods, nor to reassess the Standard NDE flaw size for commonly used method that would evaluate advances in NDE tools, processes, and equipment since the 1970s. This paper provides an overview of the "A Survey of NASA Standard Nondestructive Evaluation," [3] and NASA's first Standard NDE methodology published in "Guidebook for the Design and Analysis of a NASA Standard Nondestructive Evaluation (NDE) Probability of Detection (POD) Study (2022)" [5].

NASA STANDARD NDE HISTORY

The concept of utilizing a Standard NDE flaw size was introduced in the "Space Shuttle Program Orbiter Fracture Control Plan (1974)" [6]. The concept was carried forth in the "Fracture Control Requirements for Payloads Using the National Space Transportation System NHB 8071.1 (1988)" [7], Marshall Space Flight Center "MSFC-STD-1249 Standard NDE Guidelines and Requirement for Fracture Control Programs (1985)" [8], and eventually into the latest version of the primary NASA standard, "NASA-STD-5009B Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components (2019)" [2]. The Standard NDE flaw sizes for fluorescent penetrant, radiographic, ultrasonic, eddy current, and magnetic particle methods were linked to a series of POD test programs performed by SSP prime contractors, which were combined and jointly analyzed in Bishop (1973) [9]. Traceability to these original studies and evolutionary changes in NASA requirements were not adequately documented. It was discovered that some of the flaw sizes were based on quantitative analysis performed by Bishop, others were based on undocumented engineering judgement and additional uncited data sources.

Salkowski [10] provides a high-level overview of the fracture control motivation for assessing NDE detection reliably and identifies the foundational POD studies used in the SSP development. Salkowski explains that the reusable orbiter design drove NDE to detect smaller flaws than could be found in proof test, which was the approach used in 1966 for Apollo-era pressure vessels. The SSP Orbiter Fracture

Control Plan required the definition of reliably detectable flaw sizes for common NDE methods that are assumed to exist by fracture analysts for the purpose of crack growth analysis. These requirements drove significant advancements in POD study design and analysis. Rummel [11] provides a helpful overview of the history of nondestructive inspection reliability and highlights the initiatives by NASA and the United States Air Force (USAF) to develop NDE methods and reliability assessments in the 1970s. The SSP's introduction of the linear elastic fracture mechanics is cited as motivating the first POD data analysis procedures, which became the ubiquitous metric of NDE reliability. Forsyth [12] provides an excellent overview of the practice of assessing NDE performance and provides early references to the first discussions of nondestructive testing reliability in 1965 involving the Atomic Energy Commission, and the subsequent motivation for NDE advancements for the nuclear power industry. Forsyth also cites the concurrent 1970s initiatives of NASA and the USAF motivated by damage-tolerance philosophies in design and maintenance, where it is assumed that parts contain undetected flaws in their as-manufactured condition that could propagate under operational service conditions. Rummel [11] and Forsyth [12] set the historical context of NDE reliability assessment state-of-practice when NASA's Standard NDE flaw sizes were developed.

The First NASA Standard NDE POD Study Design and Analysis Methodology

Bishop [9] was a pioneering POD study that featured multiple inspectors from three facilities to evaluate inspector-to-inspector variability in the a90/95 flaw sizes for the common NDE methods of radiography, ultrasonic, eddy current, and penetrant. This study is the first known reference to propose an approach to estimate the flaw size that a large proportion of inspectors would reliably detect, which became known as the NASA Standard NDE flaw size. The first definition of Standard NDE flaw sizes in the SSP Orbiter Fracture Control Plan were based on the results of Bishop's study, and ultimately, NASA-STD-5009B's flaw size for penetrant and radiography can be directly traced to Bishop's results.

The Bishop POD study featured 420 fatigue cracks in 2219-T87 aluminum alloy specimens presented to 5 to 7 inspectors for each NDE method. The 420 flaws were induced in 164 specimens fabricated by two contractors, with nominal specimen dimensions of 4 inches wide and 16 inches long in thicknesses (t) of 0.060 inch and 0.210 inch. Each specimen contained multiple fatigue cracks, with some specimens having fatigue cracks on both sides. All of the flaws in this study were open surface, partly through cracks (PTCs). Bishop reports that "The location and occurrence of the flaws were carefully selected to eliminate any pattern effect which may have been detected by the inspectors." Contrary to current NDE best practices, there were no blank (unflawed) specimens included in this POD study, and therefore the probability of a false call was not estimated.

The study included flaws of varying aspect ratios, and the practice of inducing the specified crack sizes from starter notches in the specimens was reported to require significant development effort. Bending and tension-tension loading were utilized, and in some cases sequential combinations of bending and tension-tension were employed to induce fatigue cracks in the specimens. It was reported that the bending mode was employed to grow longer cracks, while tension-tension was used to grow deeper cracks. Bending cracks tended to produce more open cracks and were sometimes finished by tension-tension loading to close the cracks more tightly. Conversely, cracks grown in tension-tension may be tightly closed, and some may have been finished using a bending mode to open the cracks. These challenge in producing consistent crack specimens affects the interpretation of the POD results since some cracks may have been more detectable based on the manner in which they were induced.

After the cracks were grown and the specimens were machined to final size, they were chemically etched at two levels of material removal by the two fabricators. Rummel et al. [13] reports that about 75% of Bishop's specimens were chemically milled to remove 0.002-inch of material thickness, while Anderson et al. [14] reports an etching rate and time of exposure that suggest a material removal of about 0.0008-inch thickness for the other 25% of the cracks. Rummel reports with respect to the 0.002-inch etched specimens that "Many of the cracks were visible on close visual inspection after chemical milling." It was a significant finding that all of the inspection data published in Bishop's POD study were conducted on etched cracks, and therefore the estimated detectable flaw sizes for some methods may not be representative of inspections on as-machined (un-etched) condition. While there is a requirement for etching before penetrant inspections, it is not required for eddy current or radiography. Rummel [13] indicates that etching improved the performance of other methods, e.g., eddy current and radiographic inspections were enhanced by etching, and therefore, Bishop's results may not be representative of common practice of non-etched eddy current or radiography.

The flaw length and depth were measured by destructive analysis of the specimens after the NDE inspections were completed. From the measured flaw length and depth, the projected elliptical face area of the flaw (i.e., a thumbnail crack shape) and the flaw depth to specimen thickness ratio were computed. The detectable flaw size for radiography was reported as the ratio of flaw depth-to-thickness of specimen, and the detectable flaw size for eddy current was reported as a function of flaw depth. For penetrant and ultrasonic, the detectable flaw sizes were reported as a function of crack face area. Flaw length is currently the most common parameter for penetrant POD analysis. Bishop's rationale for using face area was:

> "For penetrant inspection, flaw detection is dependent on the visibility of the fluorescence against the test specimen background. The brightness of the indication is the controlling factor and is proportional to the amount of fluorescent material absorbed by the developer. Although the proper flaw parameter for penetrant would appear to be crack volume, crack area was used in this study because the actual crack width information was not available. Since the crack volume equals the crack area times a factor (crack opening), crack area best approximates crack volume for this study."

The inability to measure crack opening width drove Bishop's choice of flaw parameterization in the analysis, and there was not an explicit assumption that detection capability depends on the flaw area. In subsequent usage of Bishop's results, there was an extrapolation that cracks of equivalent area are equally detectable. However, that was not the original motivation. Furthermore, no POD study has been discovered that supports this general assumption of detectability based on equivalent area.

While varying crack aspect ratios were included in Bishop's dataset, the reported detectable flaw sizes were not reported as a function of aspect ratio. Furthermore, it was discovered that the distribution of flaw sizes in Bishop's study does not support modelling the detectable flaw size as a function of aspect ratio. However, in the SSP Orbiter Fracture Control Plan, the idea of detectability was extrapolated to be related to flaw aspect ratio.

The inspectors for a given NDE method were presented with all of the flaws. Each inspector reported a flaw was present (i.e., a hit), or that a flaw was not present (i.e., a miss). Recall, there were no blank (unflawed) specimens included, and therefore if an inspector does not detect the flaw, it is a miss rather than a true negative of no flaw being present. The inspectors were alphabetically labeled in the reported data, and they were not associated with their respective facility. There were 5 inspectors for ultrasonic and eddy current methods, and 7 inspectors for radiographic and penetrant. The complete dataset was transcribed from Bishop's report, and it is contained in [3].

Bishop's POD Analysis Methodology

Bishop's approach to estimating the detection capability for a large proportion of inspectors involved two stages, where the first was an estimate of each individual inspector's 90/95 detection capability (i.e., a90/95 flaw size). This individual POD analysis approach was based on a binomial distribution that assumes the probability of detecting a flaw increases with flaw size (i.e., larger flaws are more probable to detect than smaller ones). Conceptually, Bishop's approach is similar to the point estimate method (PEM) referenced in NASA-STD-5009B, derived from Rummel [15], and commonly used in NASA Special NDE demonstrations. In a PEM demonstration, 29 flaws of the same nominal size are presented to an inspector, with a number of blank (unflawed) specimens, and the inspector must find all 29 flaws to demonstrate they possess at least 90% POD of that flaw size with 95% statistical confidence, and it is often referred to as the 29/29 method. However, in Bishop's analysis, an ordered range of flaw sizes were included in grouping, rather than a single nominal flaw size in a PEM demonstration. Bishop called the analysis approach the "sorted group ascent method" (SGAM), which is described as:

> "The sorted group ascent method was devised in order to determine a subset of flaws, characterized by a flaw size, which met set values of test confidence and probability of detection. Starting at the bottom of a sorted list, the largest consecutive subset is determined which meets the … size requirement for the number of misses encountered. The smallest value of the sorted flaw size completely contained with this subset characterizes the subset and can be reasonably called the flaw sensitivity limit."

Similar to the PEM, the number of misses from the number of inspections had to meet a specified threshold to demonstrate at least 90% POD with 95% confidence. However, the flaw sizes within the Bishop group of inspections were widely variable. For example, in the analysis of an eddy current inspector, the flaw depth within the subset ranged from 0.017 inch to 0.178 inch in depth. By including a range of flaw sizes, it violates the binomial assumption that the detectability of all flaws within a subset have the same POD. Spencer [16], in regard to the eddy current example, states that the sorted group ascent method, "…is not conservative and ignores the detection rates immediately surrounding the minimum flaw size in the group, which is said to be the 90/95 flaw size."

Salkowski [10] discusses the motivation for employing the SGAM and its technical faults, as:

> "As was typical for the time, Rockwell based its statistical analysis on the binomial distribution. The binomial approach assumes that the POD for all cracks of a given size is constant. In order to obtain a reasonable confidence bound on the POD it is necessary to have a large sample of equal size cracks. Since the test panels contained only a few cracks at any given size, Rockwell was forced to increase the effective sample size by grouping together cracks of different sizes. Since different cracks cannot logically have the same POD, grouping violates the necessary assumption that the POD is constant. While the lower confidence bound derived from grouped cracks is technically not valid, it was the only method available at the time."

While this criticism of Bishop's analysis was published in 1995, there were no revised analyses of Bishop's dataset discovered during the 2022 retrospective survey. As will be discussed later, a re-analysis of Bishop's POD data was performed in 2022.

The SGAM analysis was reproduced as a part of this NESC review to evaluate its performance and additional technical issues with the methodology were discovered regarding non-unique flaw size solutions that satisfied the algorithm, and there was no explanation of how those cases were resolved in Bishop's data analysis section. Of more concern, Bishop reported that extrapolation was required to report the detectable flaw size when the specified number of hits out of a number of trials to provide 90/95 detection capability were not achieved. The need to extrapolate reflects on a weakness of the SGAM approach, and the insufficiency of the flaw size distribution to define the transition from an inspector's probability of non-detection-to-detection as a function of flaw sizes (i.e., POD curve transition region).

While SGAM appears crude compared to current analysis approaches in MIL-HBDK-1823A, Bishop's analysis represented significant advancement compared to the state-of-the-practice in 1973. SGAM was subsequently refined by others, including Rummel [13] to create smaller subgroups of flaws with less variable flaw sizes, known as the "moving average approach," that better conforms to the constant flaw size assumption. In less than 10 years after Bishop's report, these binomial point estimate methods began to be replaced by the generalized linear model regression approaches used today, where the most common is logistic regression, originally proposed by Berens and Hovey [17].

Bishop reported flaw sizes for 95% POD and 90% POD, both with 95% statistical confidence. At the time of Bishop's report, 90% POD with 95% confidence had not been agreed upon as the standard for reporting detection capability. Bishop offers the following discussion regarding a95/95 and a90/95.

> "An evaluation of the 0.90/95% and 0.95/95% limits for the data of this study are shown… It is logical to think in terms of 95% test confidence and 95% operator confidence because these figures represent the two-sigma limit, a commonly used value for process control and other quality control statistics. The probability-of-detection value should be selected near the bend in the top part of the S-curve. The larger the value at which the fraction can be chosen, the more reliable the inspection process will be and the more confidence the inspection will generate."

However, Bishop did not make a definitive recommendation as to which detection limit should be used in practice. The use of 90/95 came later, and documentation on its rationale was not discovered in this retrospective survey.

The second step in estimating the flaw size that a large proportion of inspectors are expected to detect involved computing the average and standard deviation across the individual inspector a90/95 flaw sizes. Using these statistical quantities, a Student's t-distribution was used to estimate a 95% coverage over the group of inspectors.

Regardless of the various weaknesses that have been cited throughout this review of Bishop's analysis, this study is commended for its pioneering contribution of proposing an approach to quantitatively estimate a flaw size that would be detectable by a large proportion of inspectors based on a representative sample of inspectors. Bishop's original explanation of how this quantity is to be interpreted as the 0.90/95%/95% flaw size is the basis of what ultimately became referred to as the Standard NDE flaw size.

DEVELOPMENT OF A NASA STANDARD NDE METHODOLOGY

The lessons learned from the retrospective survey of NASA Standard NDE [3] directly influenced the NASA "Guidebook for the Design and Analysis of a NASA Standard Nondestructive Evaluation (NDE) Probability of Detection (POD) Study" [5]. MIL-HDBK-1823A [4] is the industry standard of practice for planning, conducting, and analyzing POD studies and serves as the primary reference for NASA's Standard NDE guidebook, however, an approach for a standard NDE type of study is not contained in MIL-HDBK-1823A. From a NASA perspective, MIL-HDBK-1823A predominantly addresses NASA Special NDE POD demonstrations that are performed by every inspector who will conduct inspections on a specific flight component. In addition, MIL-HDBK-1823A implies the context where the detection capability is being discovered (i.e., first estimated for a specific application or new NDE method or application geometry). However, in a Standard NDE study, there is an expectation that the NDE method is mature, and its capabilities are characterized and documented based on extensive developmental efforts and prior POD studies before embarking on a resource intensive Standard NDE study. This advantageous prior knowledge is assumed and leveraged in NASA's methodology to strategically and efficiently plan a Standard NDE POD study.

Consistent with MIL-HDBK-1823A, Section 4.5.1.b, a statistician should participate in the planning stages of a Standard NDE POD study and remain involved throughout the analysis and reporting. This helps to ensure that the study is efficiently designed to meet its objectives of characterizing POD capability with sufficient precision and helps to avoid reporting potentially erroneous and misleading results. A statistician's perspective helps to identify inadvertent study design weaknesses that may restrict the POD analysis, and design efficiency may be gained by employing statistical design of experiments principles to strategically specify the required data with minimal cost. MIL-HDBK-1823A summarizes the fallacy in assuming that a statistician's primary role occurs in the data analysis by stating "Poor planning cannot be remedied after the data are collected."

Overall, the NASA Standard NDE guidebook offers valuable reminders for experienced readers and an introduction of important aspects for novice readers. Considerations of each design phase are covered, without being overly prescriptive, and it recognizes the need to accommodate unique aspects of specific applications and methods. The appendices provide Standard NDE study analysis examples, inspector sampling strategies, alternative statistical approaches, and a concise checklist of guidance for the design and analysis of a Standard NDE study. Beyond the design and execution of a Standard NDE POD study, the guidebook is expected to be a valuable resource for a fracture control analyst in critically evaluating and interpreting Standard NDE flaw sizes.

The guidebook proposed the Standard NDE study requirements succinctly as follows, which was adopted as the first requirement of this type in forthcoming release of NASA-STD-5009C.

> A Standard NDE POD study shall consist of a MIL-HDBK-1823A compliant POD study that is conducted by a minimum of 10 inspectors that form a representative sample from a specific population of inspectors. Individual inspector analyses shall be performed in accordance with MIL-HDBK-1823A methods, and the estimated a90/95 flaw sizes for the individual inspectors shall be reported. Individual inspector probability of false calls (POF) shall be reported and are recommended to not exceed 1% POF with 50% confidence. The Standard NDE flaw size shall be estimated as a function of the average and standard deviation of individual inspector a90/95 flaw sizes, and it shall represent the flaw size that 90% of inspectors are expected to demonstrate at least 90/95 detection capability. Approval of the study design, execution, and analysis, or waivers from these parameters, are subject to review and approval of the responsible Fracture Control Board.

These requirements assume extensive NDE and statistical expertise in POD to plan and conduct an acceptable NASA Standard NDE POD study, and the NASA Standard NDE guidebook is considered as a useful complementary resource to fulfil this requirement.

Standard NDE Study Design

The design of a Standard NDE POD study requires both NDE and statistical design expertise. It involves defining the NDE method and procedure, specimen characteristics, statistical flaw size design, statistical inspector sampling plan, and independent flaw characterization. The study design is guided by the intended analysis approach and application scenarios of the resultant Standard NDE flaw sizes, which often require an evaluation of similarity and transferability to specific flight components [18]. The Standard NDE POD study design decisions and their associated rationale form a significant portion of the documentation of the study, and it provides traceability of Standard NDE flaw sizes and helps to ensure the integrity and reliability of NASA's spaceflight system analyses that rely on them.

Specimen Characteristics

A Standard NDE study is expected to be broadly applicable to multiple programs, components, and inspection facilities, and therefore, it is assumed that in most cases the specimen geometry will be simple in nature (e.g., flat panels, solid round bars, or tubular cross-sections) rather than a complex flight component geometry. The Standard NDE flaw sizes in NASA-STD-5009B are routinely interpreted to apply to most aerospace metallic alloys, even though some of the POD studies were performed on a single alloy. In general, careful consideration should be given to the specimen geometry, material, and flaws to be representative or conservative relative to field inspections. Material representativeness includes consideration of the material residual stress state, surface finish, and any other conditions that might affect flaw detectability of the NDE method and are consistent with the material state of flight components. Fatigue cracks have traditionally been considered to be worst-case, conservative flaws in evaluating the POD of methods for surface imperfections in metallic components.

For Standard NDE, naturally occurring or simulated induced flaws (e.g., fatigue cracks) that provide representative flaw-to-flaw variability are recommended instead of simulated fabricated flaws (e.g., electro-discharge machining (EDM) notches). While it would be desirable to utilize naturally occurring flaws, it is assumed that there will be an insufficient number of flaws that can be independently characterized available for a Standard NDE study. Therefore, induced flaws will be used in most Standard NDE studies that are representative of flaws arising from a component's fabrication and operational usage. Induced flaws should have a defined crack morphology (e.g., aspect ratio and crack opening) that has been assessed by a materials engineer as being representative of or conservative to naturally occurring flaws. While open surface fatigue cracks are anticipated to be the most common flaw type in Standard NDE studies, other types of flaws (e.g., edge or corner cracks) may be utilized depending on the intended application of the resultant Standard NDE flaw sizes.

Statistical Flaw Size Design

The primary flaw size characteristic (e.g., depth) that will be related to POD is defined by the specific NDE method, and potential secondary flaw characteristics (e.g., aspect ratio) that may influence POD should also be considered. Statistical design parameters of the flaw size design include the flaw maximum and minimum size, the distribution of flaw sizes across the range of interest, the number of flawed and unflawed specimens/sites, and the number of replicated flaws.

Conceptually, the flaw size distribution should include a range of flaws that span from rarely detectable (POD near zero) to consistently detectable (POD approaching 1). Recommendations on the range of flaw sizes to be included in the study depend on whether the method is signal-response or hit/miss. For hit/miss NDE methods, Annis et al. [19] Section 6.2 recommends a maximum flaw size at the a97 (POD of 97%), and a minimum flaw size at the a3 (POD of 3%), which are flaws sizes that may be inferred from prior developmental POD studies. Beyond this range, extremely large flaws that are always detected and extremely small flaws that are never detected provided limited value in estimating the a90/95 flaw size.

MIL-HDBK-1823A, Section 4.5.2.2.a generally recommends uniformly spaced flaw sizes between the maximum and minimum flaw sizes in the study. However, it also suggests a concentration of flaws around the a90 region may be beneficial, and concentration of flaws in the transition region near a50, which is supported by Safizadeh et al. [20]. For hit/miss methods, there should be sufficient overlap of hits and misses in the vicinity of the a50 flaw size, which is the steeply increasing portion of a POD model that forms the transition region from misses to hits. Henry et al. [21] defines an approach to quantify the overlap as the percentage of flaws between the smallest hit and largest miss. In general, approximately 50% overlap is recommended, which means that 50% of the study's flaws are in the transition region. Complementary to the characterization of overlap, Henry [21] defines evenness as the percentage of misses in the POD study and suggests values of 30 to 50%. Considering these characteristics of overlap and evenness supports reliable, unbiased modeling of the POD model. These characteristics of the flaw design require prior knowledge of the NDE method's POD curve for a single inspector or a small group of inspectors, which is expected to be available in the design of a NASA Standard NDE study.

For signal-response methods, the maximum flaw size should be chosen to avoid saturation of the signal that occurs when a further increase in flaw size does not result in an increase in signal. As an example, for eddy current, beyond a certain depth, the signal-response saturates and no longer increases with deeper cracks. In a similar consideration, the minimum flaw size may be limited by the physics of the inspection method. The range of flaw sizes in the study should be chosen to reside within a range that avoids lower or upper saturation limits, where signal or detectability are no longer proportional to flaw size. In addition, the maximum flaw size should not greatly exceed the flaw size associated with the signal decision threshold. The minimum flaw size in the study should be below the flaw size associated with the decision threshold.

The number of flaws in a POD study depends on whether the NDE method is signal-response or hit/miss call. MIL-HDBK-1823A, Section 4.5.2.2.b recommends 40 flaws for signal-response methods and 60 flaws for hit/miss methods. These are considered a reasonable number of flaws based on typical practice. For a Standard NDE POD study, a marginal increase in the number of flaws may be considered if there is less confidence on prior POD studies. For a hit/miss method, Henry [21] suggest that 90 flaws distributed with acceptable overlap and evenness is beneficial, and there is marginal benefit beyond 90 flaws if the flaw size distribution is satisfactory to cover the range described. While the number of flaws in a Standard NDE study generally receives the most attention, the distribution of the flaw sizes is equally, if not more, important. A study that features fewer well-distributed flaws sizes over an appropriate range may be more effective in estimating the POD model than a larger study with poorly planned flaw sizes. Flaw specimen fabrication is a primary driver for the cost of a POD study, and therefore, the number of flaws will have significant impact on the overall study resource demands.

Unflawed specimens/sites are included in the study to preclude inspector guessing and to estimate the POF. NASA-STD-5019A specifies 90/95 POD, but does not specify a required POF, and the NASA-STD-5009B Standard NDE flaw sizes are not provided with an associated POF level. NASA-STD-5009B requires that POF for Special NDE be reported, but no maximum POF is specified. In the absence of requirements, the guidebook recommends 1% POF at 50% confidence. This leads to a recommendation for a minimum of 60 unflawed specimens/sites with no false positive detections for a hit/miss method, and a minimum of 40 unflawed specimens/sites are used for a signal-response method based on the Limited Sample POD methodology [22].

Counter to conventional thought, a Standard NDE study does not need to be large if it is well designed. Bishop’s large comprehensive study with 420 fatigue cracks resulted in an implicit assumption that all Standard NDE studies require a larger number of flaws, and therefore, they are extremely expensive to conduct compared to a traditional POD study. However, the NASA guidebook suggests that a more modest number of flaws for Standard NDE with multiple inspectors.

Statistical Inspector Sampling Plan

The selection strategy for the representative group of inspectors is one of the most critical and influential aspects of a Standard NDE POD. This statistical sampling plan is informed by the intended application(s) of the Standard NDE flaw sizes. The term sampling denotes that a relatively small proportion of possible inspectors are chosen from a population of inspectors will participate in the Standard NDE study. If sampled appropriately from the population, then the POD detection capability from this small sample of inspectors can be used to infer the capability of the entire population of inspectors.

The specific population of inspectors may be defined by factors including: the inspectors’ certification level, industry (e.g., aerospace), component type (e.g., pressure vessels), facility, or by a contractual arrangement related to a specific NASA program. For example, NASA-STD-5009B stipulates National Aerospace Standard NAS-410 [23] Level 2 or higher certification of inspectors in which Standard NDE flaw sizes are assumed detectable. Salkowski [10] discusses the importance of the population of inspectors within the aerospace industry that are typically seeking to detect smaller flaws than those in other industries (e.g., railway systems where larger flaws are generally of interest). The population of inspectors may be defined by the routine inspection of specific aerospace components, e.g., composite overwrap pressure vessel (COPV) metallic liners require stringent penetrant inspections and inspectors who have extensive experience with these components may form the specific population of interest. Lastly, the population may be defined by an organization (e.g., inspectors within a specific facility or contractor) or by a collection of facilities and contractors supporting a single contract or program, as was the case in the Bishop study involving the SSP prime contractors. In the usage of Standard NDE flaw sizes, the sampling strategy inherently assumes that the sample of inspectors is representative of current and future inspectors within the population of interest, if they undergo similar training, inspection experience, and certification.

Estimating inspector-to-inspector variation is the primary objective of a Standard NDE, and therefore it follows that including more inspectors will provide more information of the variability of individual inspector’s detectable flaw size and increase the precision of the Standard NDE flaw size. While flaw-to-flaw variation is typically cited as a major source of variability in POD modeling, inspector-to-inspector variability may be an equally large component of variability in a Standard NDE study.

In Bishop, between 5 and 7 inspectors per NDE method were chosen to inspect 420 flaws. In this example, there were 420 flaws and a relatively small number of inspectors. In contrast, a comprehensive study conducted by the United States Air Force (USAF) colloquially referred to as “Have Cracks, Will Travel,” described in Lewis et al. [24] featured numerous inspectors and a relatively small number of flaws. Koh and Meeker [25] explored a subset of this USAF database with 98 inspectors performing eddy current inspections on 52 flaws. Comparing Bishop to Lewis illustrates different POD design philosophies in the ratio of flaws to inspectors.

MIL-HDBK-1823A discusses random sampling of inspectors, but no guidance on the number of inspectors is provided. The Air Force's "Recommended Processes and Best Practices for Nondestructive Inspection (NDI) of Safety-of-Flight Structures," [26] recommends "…at least 10-percent of the inspector population or at least 10 inspectors be included in the experimental design, whichever is larger." Therefore, a minimum of 10 inspectors are recommended to be chosen from the representative population of interest to conduct a NASA Standard NDE study, however, accommodations for a smaller number of inspectors are discussed. In regard to POD study resource demands, the relative cost of including additional inspectors is expected to be small relative to the cost of producing the specimens and performing independent flaw size characterization.

Independent Flaw Size Characterization and Execution Protocol

A strategy for the independent characterization of the flaw sizes should be developed in the planning stage of the POD study. POD statistical modeling assumes that flaw sizes are known without error, and violating this assumption by using nominal or approximated flaw sizes in the analysis can produce misleading results. Essentially, using assumed flaw sizes in the POD analysis rather than independently measured flaw sizes introduces another component of variability that can bias the estimated a90/95 flaw sizes. Destructive flaw characterization is commonly used. Computed topography (CT) advancements may provide sufficient independent characterization nondestructively. There are clear advantages of measuring the flaw size in its as-inspected state nondestructively, and it allows for the specimens to be re-inspected in future POD studies, which maximizes the value of the specimen production investment.

The execution protocol of a Standard NDE study should be documented and independently monitored. MIL-HDBK-1823A, Section 4.5.3 recommends that a test monitor is designated to assure that guidance provided in the execution protocol is followed. A designated test monitor for the Standard NDE study should be present at each facility during the inspection process. Independent oversight of the inspection process improves the reproducibility and validity of the resultant Standard NDE flaw sizes. A documented briefing to provide consistent instructions to the inspectors and/or facilities participating in the study is recommended. Without adherence to the instruction provided, one group of inspectors may inadvertently gain an advantage in the inspection process. The USAF's "Have Cracks, Will Travel" [23] describes pre-recorded audio briefings synchronized with a slide presentation to ensure that the information shared at each facility would be identical to avoid potential bias in attitude and understanding.

Before executing the Standard NDE study, a detailed specimen physical cleaning and inspection is conducted with photographic documentation to establish a baseline condition for future reference and revalidation. Based on this inspection, a primary set of specimens is identified that excludes specimens with questionable specimen or flaw characteristics that may influence an inspector's ability to detect the presence or absence of a flaw, or positively or negatively influence the detection capability of the NDE method. A secondary set of specimens is useful for training and technique development.

Inspections of the primary specimen set used to derive the Standard NDE flaw size are to be performed in a blind manner, meaning that the inspector has no knowledge of whether it is a flawed or unflawed specimen/site nor does the inspector have an indication of the flaw location on the specimen. Each inspector is presented the flawed and unflawed specimens/sites in a unique randomized order to preclude the ability to detect a pattern that might lead to inspector familiarity or guesses regarding flaw presence. Noninformative specimen designations are assigned randomly, so that a specimen's markings and designation should not be indicative of the specimen characteristics (e.g., whether a flaw is present, its size, or location). Furthermore, if numbers are used in the specimen designation, the sequence should not be correlated with any flaw characteristic (e.g., flaw size increasing with the specimen number). In general, every reasonable effort should be made to avoid suggesting any details on the specimen characteristics to the inspector.

Inspector fatigue in conducting sequential inspections should be considered in the execution protocol. This applies to manual inspection methods and the manual review of scans produced from imaging and/or automated techniques. An acceptable inspection duration should strive to be consistent and representative of the expected inspection period of the intended operational field inspections. As a consideration, if an inspector is not time constrained during the POD study, they may tend dwell longer on a specimen than an operational inspection of a flight component, and this may result in a better detection capability in the laboratory that is not representative of operational inspection capability. In addition, an attempt to mimic the physical posture of the inspector during field inspections may be considered, and this may contribute to the inspection duration consideration.

Standard NDE Statistical Analysis

The analysis of NASA Standard NDE study is a two-step process that begins with analyzing individual inspector detection capability and is followed by an analysis of variability in the individual inspector a90/95 detectable flaw sizes. Analyzing individual inspector detection capability first leverages an NDE engineer's familiarity with traditional POD modeling and the ability to utilize available software tools. It enables a more intuitive and insightful review of the individual inspector a90/95 flaw sizes. Bishop's seminal study used a conceptually similar two-step approach, so it is also historically consistent.

NASA-STD-5019A requires that the NDE detectable flaw size has 90% probability of detection with 95% confidence, but there is no NASA requirement for the proportion of inspectors that will possess this detection capability. To quantitatively define a proportion of inspectors in the absence of an existing NASA requirement, rationale was developed based on historical precedence and experience, in consultation with NDE engineers, statisticians, and the fracture control community. It is recommended that 90% inspector coverage at 50% confidence based on individual inspector a90/95 flaw sizes is reported, however, the methodology presented can be adapted to other choices of inspector coverage probability.

A significant portion of the guidebook was devoted to the planning of a Standard NDE study because careful attention to the study design and execution protocol are critical to support statistically defensible analyses. It is a common misconception that the study design and analysis aspects are independent of analysis methods, which often leads to inadvertent analysis limitations after the specimens are fabricated and the inspections are completed.

Estimating Individual Inspector a90/95

A Standard NDE study can be thought of as a collection of individual inspector POD studies, where every inspector sees every specimen. While it is not strictly required for statistical modeling that every inspector sees every specimen, it simplifies the Standard NDE analysis approach, and it is expected to be common in practice. The inspection results from an individual inspector are analyzed using MIL-HDBK-1823A methods to estimate an individual inspector's a90/95 flaw size. While the analysis approaches of MIL-HDBK-1823A are expected to be sufficient in most cases, it is acknowledged that more complex statistical models may be required, and consultation with a statistician is recommended.

Estimating Individual Inspector Probability of a False Positive

The individual inspector calls from unflawed specimens or sites are analyzed to estimate an individual inspector's POF. This analysis of unflawed specimens/sites is referred to as a noise analysis in MIL-HDBK-1823A. If an inspector calls a hit at a specimen location that does not contain a flaw, then it is a false positive indication. False positives can occur in signal-response and hit/miss NDE methods, and there are distinctive analysis approaches for each method type.

For a signal-response method, the signal average and standard deviation from the collection of unflawed specimens/sites are computed, and a one-sided statistical tolerance interval estimates the upper bound on the signal associated with 99th percentile (i.e., 1% upper quantile) with 50% confidence. If the 1/50 signal level falls below the signal decision threshold of the NDE method, then a maximum of 1/50 POF is successfully demonstrated. For a hit/miss method, the number of false positive indications is recorded and compared to the total number of unflawed specimens/sites inspected. If 60 unflawed specimens/sites are included in the study, then no false positives are allowed to demonstrate 1/50 POF. Individual inspectors in a Standard NDE study are expected to demonstrate a maximum of 1/50 POF. If an inspector exceeds 1/50 POF, a diagnostic phase is entered to identify a correctable cause.

Estimating Standard NDE Flaw Size

The individual inspector a90/95 flaw sizes are used to estimate the Standard NDE flaw size in the second step of the analysis. The average inspector a90/95 flaw size and the standard deviation of the a90/95 flaw sizes across the inspectors are used to estimate a statistical tolerance interval (i.e., a one-sided confidence bound on a quantile) that represents the proportion of the inspector population expected to demonstrate at least 90/95 detection of the Standard NDE flaw size. This approach is straightforward to implement, and its simplicity enhances the insights regarding the range of detection capability of individual inspectors sampled from the population of interest.

Standard NDE Analysis Methodology Applied to Historical Data

To baseline the Standard NDE analysis methodology, it was applied to Bishop's POD study. This activity also provided a quantitative assessment on NASA's current Standard NDE flaw sizes in NASA-STD-5009B. Analyses were performed for all four NDE methods adopting Bishop's flaw parameterizations of a/t for radiography, flaw face area for penetrant and ultrasonic, and flaw depth for eddy current and is documented in [3]. As described, individual inspector average and standard deviation of a90/95 flaw sizes were computed and the 95% coverage with 50% confidence was used to estimate the a90/95/95 flaw size. Note that 95% coverage was used in this retrospective analysis to be consistent with Bishop's analysis, however, 90% coverage is now recommended in the NASA Standard NDE guidebook. As an example of the results, Table 1 provides the individual a90/95 detectable flaw sizes for each inspector and the Standard NDE flaw size for radiography, where POD is considered a function of flaw depth (*d*) to specimen thickness (*t*), expressed as a percentage. A *d/t* flaw size of 100% represents a through crack.

Table 1: Radiography Detectable Flaw Sizes

Radiography	[Flaw depth (d)] / [Specimen thickness (t)] x 100		
	Inspector	**Bishop (1973) [9]**	**Parker (2022) [3]**
		(d/t %)	(d/t %)
Individual Inspector a90/95 Flaw Sizes	A	53	59
	B	64	66
	C	66	67
	D	59	70
	E	65	79
	F	60	77
	G	58	74
Inspector Average a90/95		61	70
NASA Standard NDE a90/95/95 Flaw Size		**70**	**83**

For radiography, Bishop's individual inspector a90/95 estimates were found to be generally non-conservative. This result is not surprising based on the earlier discussion of Bishop's sorted group ascent analysis methodology, and there was a similar finding for the other NDE methods. Therefore, it followed that Bishop's Standard NDE flaw sizes (i.e., the a90/95/95) were generally non-conservative, with the exception of penetrant, which fortuitously benefited from a miscalculation of square root made in 1973 in the Bishop report discovered in the 2022 survey.

CONCLUSIONS

A retrospective review of NASA Standard NDE traced the evolution Standard NDE flaw sizes from studies in the 1970s in support of the SSP to NASA current standard (i.e., NASA-STD-5009B). The lineage of NASA Standard NDE spans nearly 50 years, and the Standard NDE flaw sizes have served numerous NASA programs without a known, attributable failure due to the application of Standard NDE flaw sizes in fracture control plans. This comprehensive review provided numerous insights, lessons learned, and supported the guidance for developing a Standard NDE guidebook for the design and analysis of a NASA Standard NDE POD Study. The guidebook is the first documented methodology to conduct a NASA Standard NDE study and it enables updating the Standard NDE flaw sizes for common NDE methods and new methods. The methodology in the guidebook extends MIL-HDBK-1823A's guidance to a Standard NDE study. The proposed methodology was developed to be straightforward, intuitive, and approachable to NDE practitioners and fracture analysts to broaden its potential application. The 50-year retrospective survey of Standard NDE and the documented methodology for conducting Standard NDE studies addresses a significant, long-standing gap in the NASA NDE body of knowledge, and it supports the continued usage of Standard NDE flaw sizes in the majority of NASA's spaceflight system designs.

REFERENCES

[1] NASA Standard 5019A (2019): "Fracture Control Requirements for Spaceflight Hardware."

[2] NASA-STD-5009B (2019): "Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components."

[3] Parker, P.; Koshti, A.; Forsyth, D.; Suits, M.; Walker, J.; and Prosser, W. (2022): "A Survey of NASA Standard Nondestructive Evaluation (NDE)," NASA TM-20220013820.

[4] MIL-HDBK-1823A (2009): "Nondestructive Evaluation System Reliability Assessment."

[5] Parker, P.; Koshti, A.; Forsyth, D.; Suits, M.; Walker, J.; and Prosser, W. (2022): "Guidebook for the Design and Analysis of a NASA Standard Nondestructive Evaluation (NDE) Probability of Detection (POD) Study," NASA TM-20220013822.

[6] King, J. P.; and Johnson, K. R. (1974): "Space Shuttle Orbiter Fracture Control Plan," Space Division Rockwell International SD73-SH-0082A.

[7] NASA (1988) "Fracture Control Requirements for Payloads Using the National Space Transportation System (NSTS)," National Aeronautics and Space Administration, NHB8071.1.

[8] MSFC-STD-1249 (1985): "Standard NDE Guidelines and Requirement for Fracture Control Programs."

[9] Bishop, C. R. (1973): "Nondestructive Evaluation of Fatigue Cracks," Space Division Rockwell International, SD 73-SH-0219.

[10] Salkowski, C. (1995): "Nondestructive Inspection Reliability Assumptions for Critical Aerospace Components," *Proceedings SPIE 2455, Nondestructive Evaluation of Aging Aircraft, Airport, Aerospace Hardware, and Materials*.

[11] Rummel, W. D. (2010): "Nondestructive Inspection Reliability – History, Status, and Future Path," *18th World Conference on Nondestructive Testing*, 16-20 April 2010, Durban, South Africa.

[12] Forsyth, D. S. (2018): "Experiences in Practicing the Assessment of Nondestructive Testing Performance," *44th Annual Review of Progress in Quantitative Nondestructive Evaluation,* Volume 37, AIP Conference Proceedings.

[13] Rummel, W. D.; Todd, P. H.; Frecska, S. A.; and Rathke, R. A. (1974): "The Detection of Fatigue Cracks by Nondestructive Testing Methods," NASA CR-2369.

[14] Anderson, R. T.; DeLacy, T. J.; and Stewart, R. C. (1973): "Detection of Fatigue Cracks by Nondestructive Testing Methods," General Dynamics, Convair Division, GDCA-DGB73-02.

[15] Rummel, W. D. (1982): "Recommended Practice for Demonstration of Nondestructive Evaluation (NDE) Reliability on Aircraft Production Parts," *Materials Evaluation*, 40.

[16] Spencer, F. W. (2020b): "Response to questions/comments on Recommendations on Inspector Numbers for a Standard Methodology of NDE Characterization," informal correspondence.

[17] Berens and Hovey (1981): "Evaluation of NDE Reliability Characterization," University of Dayton Research Institute, AFWAL-TR-81-4160.

[18,]Koshti, A.; Parker, P.; Forsyth, D.; Suits, M.; Walker, J.; and Prosser, W. (2022): "Guidebook for Assessing Similarity and Implementing Empirical Transfer Functions for Probability of Detection (POD) Demonstrations for Signal Based Nondestructive Evaluation (NDE) Methods," NASA TM-20220003648.

[19] Annis, C.; and Gandossi, L. (2012): "Influence of Sample Size and Other Factors on Hit/Miss Probability of Detection Curves," EBIQ report no. 47, EUR - Materials Evaluation Scientific and Technical Research Series.

[20] Safizadeh, M. S.; Forsyth, D. S.; and Fahr, A. (2004): "The Effect of Flaw Size Distribution on the Estimation of POD," *Insight*, 46, 6.

[21] Henry, C. E.; and Kabban, C. S. (2022): "Modern Design and Analysis for Hit/Miss Probability of Detection Studies using Profile Likelihood Ratio Confidence Intervals," *Materials Evaluation*.

[22] Koshti, A.; Parker, P.; Forsyth, D.; Suits, M.; Walker, J.; and Prosser, W. (2021): "Guidebook for Limited Sample Probability of Detection (LS-POD) Demonstration for Signal-Response Nondestructive (NDE) Methods," NASA TM-20210018515.

[23] NAS 410 (2020): "NAS Certification and Qualification of Nondestructive Test Personnel," 5th Edition.

[24] Lewis, W. H.; Sproat, W. H.; Dodd, B. D.; and Hamilton, J. M. (1978): *Reliability of Nondestructive Inspections,* United States Air Force contractor report SA-ALC/MME 76-6-38-1, prepared by The Lockheed-Georgia Company.

[25] Koh, Y.M., and W.Q. Meeker (2017), "Quantile POD for Nondestructive Evaluation with Hit-Miss Data," *Research in Nondestructive Evaluation*, 30:2, pg. 89-111.

[26] Brausch, J.; Butkus, L.; Campbell, D.; Mullis, T.; and Paulk, M. (2008): "Recommended Processes and Best Practices for Nondestructive Inspection (NDI) of Safety-of-Flight Structures," AFRL-RX-WP-TR-2008-4373, Materials Integrity Branch System Support Division.