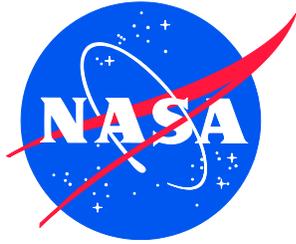


NASA/TM-20220013822  
NESC-TI-21-01657



# Guidebook for the Design and Analysis of a NASA Standard Nondestructive Evaluation (NDE) Probability of Detection (POD) Study

*Peter A. Parker  
Langley Research Center, Hampton, Virginia*

*Ajay Koshti  
Johnson Space Center, Houston, Texas*

*David S. Forsyth  
NDTAnalysis, St John, U.S. Virgin Islands*

*Michael W. Suits, James L. Walker  
Marshall Space Flight Center, Huntsville, Alabama*

*William H. Prosser/NESC  
Langley Research Center, Hampton, Virginia*

## NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.

- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- **TECHNICAL TRANSLATION.** English-language translations of foreign scientific and technical material pertinent to NASA's mission.

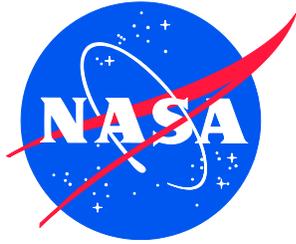
Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at <http://www.sti.nasa.gov>
- E-mail your question to [help@sti.nasa.gov](mailto:help@sti.nasa.gov)
- Phone the NASA STI Information Desk at 757-864-9658

- Write to:  
NASA STI Information Desk  
Mail Stop 148  
NASA Langley Research Center  
Hampton, VA 23681-2199

NASA/TM-20220013822  
NESC-TI-21-01657



# Guidebook for the Design and Analysis of a NASA Standard Nondestructive Evaluation (NDE) Probability of Detection (POD) Study

*Peter A. Parker*  
*Langley Research Center, Hampton, Virginia*

*Ajay Koshti*  
*Johnson Space Center, Houston, Texas*

*David S. Forsyth*  
*NDTAnalysis, St John, U.S. Virgin Islands*

*Michael W. Suits, James L. Walker*  
*Marshall Space Flight Center, Huntsville, Alabama*

*William H. Prosser/NESC*  
*Langley Research Center, Hampton, Virginia*

National Aeronautics and  
Space Administration

Langley Research Center  
Hampton, Virginia 23681-2199

September 2022

The use of trademarks or names of manufacturers in the report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148  
NASA Langley Research Center  
Hampton, VA 23681-2199  
Fax: 757-864-6500

## **Preface**

**Purpose** – This document provides guidance on the design and analysis of a NASA Standard nondestructive evaluation (NDE) probability of detection (POD) study. A Standard NDE flaw size is considered to be conservative such that most inspectors, trained and certified in the specific NDE method, are expected to provide at least 90/95 POD for that flaw size to inspect fracture-critical hardware.

**Scope** – This document is specifically applicable to NASA Technical Standards that establish the NDE requirements for any NASA system or component, flight or ground, where fracture control and a quantitative demonstration of POD is a requirement, including NASA-STD-5009B, Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components, and NASA-STD-5019A, Fracture Control Requirements for Spaceflight Hardware.

## Table of Contents

Preface.....	iii
<b>1.0 Introduction.....</b>	<b>1</b>
<b>2.0 Requirements for a NASA Standard NDE POD Study.....</b>	<b>2</b>
<b>3.0 Standard NDE Study Design .....</b>	<b>3</b>
3.1 NDE Method Specifications .....	5
3.2 Specimen Characteristics .....	6
3.3 Statistical Flaw Size Design.....	7
3.4 Statistical Inspector Sampling Plan .....	12
3.5 Independent Flaw Size Characterization.....	14
<b>4.0 Execution .....</b>	<b>15</b>
<b>5.0 Analysis .....</b>	<b>17</b>
5.1 Estimating Individual Inspector $a_{90/95}$ .....	19
5.2 Estimating Individual Inspector Probability of a False Positive .....	19
5.3 Estimating Standard NDE Flaw Size .....	20
<b>6.0 Documentation .....</b>	<b>21</b>
<b>7.0 Conclusions.....</b>	<b>23</b>
<b>8.0 References.....</b>	<b>24</b>
<b>Appendix A – Tabulated <math>k_1</math> Values to Estimate Standard NDE Flaw Size.....</b>	<b>26</b>
<b>Appendix B – Standard NDE POD Study Analysis Examples .....</b>	<b>27</b>
<b>Appendix C – Inspector Sampling Discussion and Multiple Facility Guidance .....</b>	<b>38</b>
<b>Appendix D – Alternative Statistical Approaches to Estimate the Standard NDE Flaw     Size.....</b>	<b>40</b>
<b>Appendix E – Checklist of Guidance for the Design and Analysis of a Standard NDE     Study.....</b>	<b>42</b>

## List of Figures

Figure 6.0-1. Example Standard NDE study condensed specification template. ....	23
Figure B.1. Inspection calls from Inspector 1 for hit/miss standard NDE POD example. ....	29
Figure B.2. Individual inspector POD models with $a_{90}$ and 90/95 estimates for hit/miss standard NDE POD example. ....	30
Figure B.3. 90% POD region of individual inspector POD models with $a_{90}$ and 90/95 estimates for hit/miss standard NDE POD example. ....	31
Figure B.4. Individual inspector signal versus flaw size models for a signal-response Standard NDE POD example.....	36
Figure C.1. Flow chart of inspector and facility sampling guidance. ....	38

## List of Tables

Table 3.0-1.	Standard NDE POD Study Design Considerations .....	4
Table B.1.	Hit/Miss Standard NDE POD Example Dataset with 60 Flaws and 10 Inspectors .....	28
Table B.2.	Individual Inspector Logistic Model Estimated Parameters, $a_{90}$ , $a_{90/95}$ , and Standard NDE flaw size ( $a_{90/95/90}$ ) Flaw Size for Hit/Miss NDE Method Example .....	32
Table B.3.	Signal-Response Standard NDE POD Example Dataset with 40 Flaws and 10 Inspectors .....	34
Table B.4.	Signals from 40 Unflawed Sites to Estimate POF .....	35
Table B.5.	Estimated 1/50 POF Signal for Each Inspector .....	35
Table B.6.	Individual Inspector Model Estimated Parameters, $a_{90/95}$ flaw size, and the Standard NDE flaw size, $a_{90/95/90}$ , for the Signal-Response NDE Method Example .....	37

## Nomenclature

2c	Flaw Length
2c/a	Aspect Ratio
a	Flaw Depth
a50	Estimated flaw size at 50% POD
a90	Estimated flaw size at 90% POD
a90/95	Estimated flaw size that provides 90% POD with 95% confidence
aXX	Estimated flaw size at XX% POD
c	1/2 Flaw Length of open surface flaw
COPV	Composite Overwrap Pressure Vessel
CT	Computed Topography
EDM	Electro-Discharge Machining
Gauge R&R	Gauge Repeatability and Reproducibility
GLMM	Generalized Linear Mixed Model
LS-POD	Limited Sample POD
MIL-HDBK	Military Handbook
NAS	National Aerospace Standard
NDE	Nondestructive Evaluation
PEM	Point Estimate Method
POD	Probability of Detection
POF	Probability of False Calls
RMS	Root Mean Square
SGAM	Sorted Group Ascent Method
SSP	Space Shuttle Program
t	Specimen Thickness
TM	Technical Memorandum
USAF	United States Air Force

## **Definitions (adapted from NASA-STD-5009B (2019))**

**Cracks or Crack-Like Flaws:** A discontinuity assumed to behave like a crack for assessment of material or structural integrity. Referred to as induced flaws, whether naturally occurring or laboratory simulated.

**Flaw:** An imperfection or discontinuity that may be detectable by nondestructive testing and is not necessarily rejectable. Examples of flaws in metallic articles include cracks, deep scratches and sharp notches that behave like cracks, material inclusions, forging laps, welding incomplete fusion, penetration, and slag or porosity with a crack-like tail. For additive manufactured metallics, skipped layers, thermal or stress induced cracks, or inclusions are examples.

**Fully Crossed Design:** An allocation of flaw specimens such that each inspector is presented the same set of flaw specimens.

**Hit-Miss NDE Data:** Data resulting from an NDE inspection where only the determination of whether an indication is present or not is recorded. Thus, the data at each measurement point corresponds to either a yes or no, or is sometimes represented numerically as a 1 (i.e., indication present) or 0 (i.e., no indication). No signal measurements from any NDE sensor output are recorded.

**Instrument Calibration:** Comparison of an instrument response with, or adjustment of an instrument response to, known references often traceable to the National Institute of Standards and Technology (NIST). This is usually performed periodically, typically at a 1-year interval. After completing calibration, a calibration sticker with calibration expiration date is affixed to the instrument.

**Instrument Standardization:** Adjustment of an NDE instrument response using an appropriate reference standard with known size discontinuities such as electro-discharged machined slots and flat bottom holes, to obtain or establish a known and reproducible response. This is usually done prior to an examination but can be carried out anytime there is concern about the examination or instrument response. It is also commonly known as calibration prior to initiating an NDE procedure. Instrument standardization should be carried out using a minimum of three data points demonstrating expected correlation between signal response and discontinuity size.

**Naturally Occurring Flaw:** A flaw that is present in a component as a result of the normally occurring manufacturing processes or usage of the component.

**Nondestructive Evaluation (NDE), Nondestructive Inspection (NDI), Nondestructive Testing (NDT):** The development and application of technical methods to examine materials or components in ways that do not impair future usefulness and serviceability in order to detect, locate, measure, and evaluate flaws; to assess integrity, properties, and composition; and to measure geometrical characteristics.

**NDE Procedure:** A written plan providing detailed information on ‘how-to’ perform a hardware-specific inspection.

**NDE Simulated Fabricated Flaw:** A flaw that is intentionally placed in a component for the purpose of generating an NDE signal response. These can be produced by a variety of material removal processes (e.g., cutting, drilling, electrical discharge machining

(EDM), laser notching, plasma focused ion beam (PFIB) notching, etc.) or additive material forming processes.

**NDE Simulated Induced Flaw:** A flaw that is intentionally placed in a component for the purpose of generating an NDE signal response. NDE simulated induced flaws are produced by intentional loading (e.g., thermal, mechanical, etc.) to induce damage (e.g., cracks, delaminations, disbonds, etc.).

**NDE Transfer Function:** A function that describes the relationship between signal responses for an NDE method as a function of flaw size for different types of flaws (e.g., naturally occurring flaws, load induced or material removal NDE simulated flaws) or for flaws in different types of components (e.g., simple geometries such as cylinders or flat plates or structural component of interest with complex geometry).

**Signal-Response NDE Data:** Data from an inspection where the NDE sensor produces a signal output (e.g., voltage, current, etc.) that is measured and proportional to flaw size. The determination for whether an indication is present is typically made based on a threshold value of the signal response.

**Special NDE:** Nondestructive inspections of fracture-critical hardware that are capable of detecting cracks or crack-like flaws smaller than those assumed detectable by Standard NDE or do not conform to the requirements for Standard NDE as set forth in NASA Standard 5009B. Special NDE methods are not limited to fluorescent penetrant, radiography, ultrasonic, eddy current, and magnetic particle.

**Standard NDE:** NDE methods of metallic materials for which a statistically based flaw detection capability has been established. Standard NDE methods addressed by NASA Standard 5009B are limited to the fluorescent penetrant, radiographic, ultrasonic, eddy current, and magnetic particle methods employing techniques with established capabilities.

**Similarity:** The outcome of an assessment that the same POD is expected in different NDE inspection situations that might include variations in NDE method/procedure, components being inspected, and/or inspection conditions.

## 1.0 Introduction

A 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 (2016) requirements. The tabulated Standard NDE flaw sizes in NASA-STD-5009B (2019) 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 does not require individual inspector probability of detection (POD) demonstrations, which can be resource intensive.

While Standard NDE POD studies have been conducted at NASA, with the most prominent ones occurring during the development of the Space Shuttle Program (SSP) in the early 1970s, there is no NASA Standard NDE POD methodology available for adding NDE methods or updating flaw sizes for the current methods covered in NASA-STD-5009B. The methodology proposed in this guidebook builds on the lessons learned from the Standard NDE POD studies, and relies on the industry best practices contained in MIL-HDBK-1823A (2009).

A Standard NDE POD study involves multiple inspectors presented with a collection of specimens that contain flaws varying in size and inspection regions where no flaws exist. The inspectors independently report their inspection findings for each specimen, which may be in the form of a signal response that exceeds a decision threshold or an inspector's visual review of the component, an image, or a scan to call a flaw indication. From the experimental data, individual inspector a90/95 flaw sizes are estimated using MIL-HDBK-1823A POD modeling methods. The Standard NDE flaw size is estimated using the average and variability of individual inspector a90/95 flaw sizes, and it represents the flaw size that a large proportion of inspectors will detect (i.e., the largest flaw they may miss). Standard NDE POD studies are designed to be representative or conservative relative to operational field inspections where they will be applied.

MIL-HDBK-1823A (2009) is the industry standard of practice for planning, conducting, and analyzing POD studies and serves as the primary reference for this guidebook. However, an approach for NASA Standard NDE is not contained in MIL-HDBK-1823A. From a NASA perspective, MIL-HDBK-1823A predominantly addresses scenarios that NASA-STD-5009B denotes as Special NDE POD demonstrations by every inspector who will perform inspections on a specific flight component. Nevertheless, much of the guidance provided in MIL-HDBK-1823A is relevant and valuable in the planning, execution, and analysis of NASA Standard NDE POD studies. This guidebook is intended to be a consistent extension of MIL-HDBK-1823A concepts, and it is assumed that the reader is experienced in the practice of this handbook's methods.

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). However, in a NASA Standard NDE study, there is an expectation that the NDE method is mature, and its capabilities are characterized and documented based on developmental efforts and prior POD studies before embarking on a resource intensive Standard NDE study. This advantageous prior knowledge is assumed and leveraged in this guidebook to strategically and efficiently plan a Standard NDE POD study, and it is an important distinction when considering MIL-HDBK-1823A's guidance, which tends to assume less prior knowledge.

Consistent with MIL-HDBK-1823A (2009), Section 4.5.1.b, it is recommended that a statistician 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 analysis, as shown in the NASA/TM–20220013820 (Parker, et al. (2022)). 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.*” The level of statistical detail provided in this guidebook assumes that a statistician is included on the POD study team, and they are familiar with the statistical concepts and modeling approaches discussed in MIL-HDBK-1823A.

A retrospective survey of NASA Standard NDE is presented in NASA/TM–20220013820 (Parker, et al. (2022)) that traces the evolution of the Standard NDE flaw sizes in NASA-STD-5009B. The lessons learned from this comprehensive survey have directly influenced the guidance contained herein. In particular, Bishop (1973) is the seminal POD study that supported the original development of NASA Standard NDE. Numerous footnotes in this guidebook highlight features of Bishop’s POD study relative to current guidance, and they serve as illustrations of the concepts being discussed. These footnotes highlight where there are differences from the historical precedence of Bishop (1973).

A proposed Standard NDE requirement is provided in Section 2 for inclusion in NASA-STD-5009C that assumes extensive NDE and statistical expertise in POD to plan and conduct an acceptable NASA Standard NDE POD study. The remainder of this TM provides guidance on the design, execution, analysis, and documentation of a Standard NDE POD study to assist in satisfying the proposed requirement. It offers valuable reminders for experienced readers and an introduction of important aspects for novice readers. Considerations of each design phase are highlighted, without being overly prescriptive, and recognize the need to accommodate unique aspects of specific applications and methods. The appendices provide Standard NDE study analysis examples, inspector sampling strategy, alternative statistical approaches, and a concise checklist of guidance for the design and analysis of a Standard NDE study. The guidance provided herein may assist a Fracture Control Board in critically evaluating a proposed Standard NDE POD study before its execution. This guidebook is organized around the design, execution, analysis, and documentation phases, and it highlights the interdependencies among these phases in a unified perspective for a Standard NDE POD study.

## **2.0 Requirements for a NASA Standard NDE POD Study**

The expectations of a NASA Standard NDE POD study are encapsulated in the following proposed requirement and comment for inclusion in 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.

MIL-HDBK-1823A (2009) provides guidance for the design, execution, and analysis of the individual inspector POD studies conducted within a NASA Standard NDE study. Additional guidance that augments and adapts MIL-HDBK-1823A for Standard NDE is contained in this guidebook, and it exploits the advantage of prior POD studies in planning a Standard NDE POD study. It addresses an approach to specify the inspector population and inspector sampling strategy. This guidebook recommends an approach to estimate the Standard NDE flaw size that provides 90% coverage of the specific inspector population.

### **3.0 Standard NDE Study Design**

The design of a Standard NDE POD study integrates NDE and statistical design expertise to support the analysis and estimation of Standard NDE flaw sizes. It involves defining the NDE method and procedure, specimen characteristics, statistical flaw size design, statistical inspector sampling plan, and independent flaw characterization. Table 3.0-1 summarizes these 5 categories of design considerations with illustrative examples for design elements within each category and the corresponding sections in this guidebook that discuss them.

It is recommended that the design is guided by the intended application(s) scenarios of the resultant Standard NDE flaw sizes, which are anticipated to require a subsequent evaluation of similarity and transferability to specific flight components described in NASA/TM-20220003648 (Koshti et al. (2022)). While it is challenging to envision future usage of the study results over a long time span, a disciplined effort to define the scope enables a strategic and resource efficient POD study design. In addition, the sufficiency of the design can be evaluated before execution by its ability to satisfy envisioned usage scenarios. It has been observed in practice that without a clear scope of envisioned usage, a POD study may strive to be too comprehensive and become excessively resource intensive.

In the following sections, each of these 5 design categories are discussed primarily from a statistical design and analysis perspective. However, it is recognized that many of these elements are dependent on specific aspects of the NDE method and will rely on the experience of a cognizant NDE engineer designing the study. Wherever possible, design considerations are discussed in general in this guidebook without an attempt to address the specifics of current and future NDE methods.

MIL-HDBK-1823A (2009) provides helpful discussions and guidance on POD study design, and is referenced extensively in the following sections. This guidebook extends the guidance of MIL-HDBK-1823A and highlights the unique aspects that are involved in a NASA Standard NDE POD study.

The Standard NDE POD study design decisions and their associated rationale form a significant portion of the documentation of the study. Detailed documentation is vital for appropriate application of the Standard NDE flaw sizes, more specifically to evaluate similarity and transferability to specific components. This documentation 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.

Table 3.0-1. Standard NDE POD Study Design Considerations

Design Phase	Design Element	Comments and examples to clarify elements	Section Ref.
<b>NDE Method Specifications</b>	NDE method	ultrasonics, eddy current, penetrant, radiography, magnetic particle	3.1
	Instrumentation system specifications	model, software version	
	Sensor / Probe	eddy current probe specifications	
	Inspection materials	chemicals, cleaning, particle size, penetrant, developer	
	Scanning method	manual or automated, rate and spatial resolution	
	Standardization/calibration method	setup of the instrument sensitivity	
	Recording method of inspection results	manual, automated	
	Decision threshold	cite the method of determination and/or source	
	Inspection data recorded	signal-response, hit/miss, images, scans	
<b>Specimen Characteristics</b>	Specimen material and properties	aluminum, steel, annealed, heat-treat, residual stress state	3.2
	Specimen geometry	flat, round, tubular	
	Specimen size	width, length, diameter, thickness	
	Specimen fabrication processes	mill, lathe, EDM, conventional, additive	
	Specimen condition	surface finish	
	Flaw type	induced fatigue crack, edge or corner crack	
	Flaw size	length, depth, area, volume, aspect ratio	
	Flaw production	bending, tension-tension, cycles and loads, thermal	
	Flaw location	random, multiple per specimen, both sides, edge	
	Flaw orientation	random, transverse, relative to the grain	
	Final flaw fabrication processes	machining performed to remove starter notch	
	Flaw condition at inspection	as-machined, after-etch, after-proof loading	
	Etching of flaw	chemical, process, material removal specification	
<b>Statistical Flaw Size Design</b>	Prior knowledge	existing POD references, estimates of a90/95, inspector variability	3.3
	Primary flaw parameter	length, depth, area	
	Secondary flaw parameter	aspect ratio	
	Range of flaw sizes	maximum and minimum flaw size	
	Number of flaws	number of flaws in primary and training set	
	Distribution of flaw sizes	uniform, proportion in transition region	
	Number of unique flaw sizes	nominal flaw sizes specified to be produced	
	Number of replicated flaw sizes	multiple flaws of the same nominal size	
	Number of unflawed (blank) specimens	blank specimens, designated unflawed regions	
<b>Statistical Inspector Sampling Plan</b>	Population of inspectors	within a facility, across facilities, by application, by program	3.4 Appendix C
	Sampling of inspectors	random sampling, weighted sampling	
	Inspector certification/qualification	NAS-410 Level II	
	Number of inspectors	total number of inspectors and within each facility/organization	
<b>Independent Flaw Size Characterization</b>	Flaw measurements	length, depth, aspect ratio, shape, crack opening	3.5
	Method to estimate the flaw size	destructive, computed topography	
	Uncertainty in flaw size measurements	measurement tolerance, features that are inferred	

### 3.1 NDE Method Specifications

It is recommended that a Standard NDE study is specifically designed for a single NDE method, since the expected detection capability and physics of the inspection method determines the flaw specimen characteristics and flaw size design.<sup>1</sup> Several NDE methods with similar detection capability could use the same specimens. However, their POD may be based on different flaw characteristics<sup>2</sup> and different conditions of the flaw (e.g., etched versus unetched condition). It is expected that the NDE method specifications will primarily rely on the expertise of an NDE engineer.

Standardization protocols, sometimes referred to as calibration procedures, are vital aspects in a Standard NDE study that involves multiple facilities and inspectors to ensure the same system sensitivity. MIL-HDBK-1823A (2009), Section 4.3 provides guidance on calibration and states that "... the statistical POD analysis is only as good as the data on which it is based, and the data are only as good as the system that produced it, and that depends on effective calibration. (An excellent system, poorly calibrated, produces data of no consequence.)" In some cases, the decision threshold for signal-response NDE methods is determined during the standardization/calibration process using representative quality indicators or calibration blocks with reference notches. The calibration block's material residual stress state, surface finish, and any other conditions that might affect flaw detectability of the NDE method are important considerations. It is recommended that there is a consistent decision threshold utilized for inspections in a Standard NDE study. There may be situations when each inspection facility is allowed to set thresholds based on their respective internal processes (e.g., if there are proprietary techniques employed by different facilities). However, this customization is not desirable nor recommended.

The level of specific NDE method implementation controls should be based on the intended program context where the resultant Standard flaw sizes will be employed.<sup>3</sup> Whether a variety of implementations is a beneficial feature or a weakness of the Standard NDE study depends on the operational field inspection context. For example, if a flight component is specified to undergo eddy current inspection that may occur at multiple, unknown inspection facilities, it may not be possible to use the same method implementation protocol in operational field inspections. In this case, requiring these facilities to artificially employ the same protocol in the Standard NDE study would not represent the operational inspection environment.

Explicitly specifying the data-recording protocol in a Standard NDE study is recommended to support the study's traceability and reproducibility. It is recommended that the most raw form of the inspection signals are recorded and archived rather than only recording an inspector's final determination of whether a flaw is present.<sup>4</sup> Having these raw data with the inspector's call

---

<sup>1</sup> NASA/TM-20220013820 (Parker, et al. (2022)) suggests that Bishop's POD study was designed to cover a wide range of flaw sizes for radiography, penetrant, ultrasonics, and eddy current, but it was not tailored for these methods individually.

<sup>2</sup> In Bishop (1973), penetrant and eddy current detection capability focused on a similar subset of the flaws. However, POD of penetrant was modeled as a function of projected elliptical flaw area and eddy current was modeled as a function of flaw depth.

<sup>3</sup> In Bishop (1973), it appears that the eddy current method's implementation at the 3 facilities was not controlled to be identical. Some facilities used manual scanning and recording of hit/miss calls and others employed an automated scanning and recording of signals.

<sup>4</sup> In Bishop (1973), the eddy current signal was not reported or used in the POD analysis, only the final hit/miss call. Without the signals, a MIL-HDBK-1823A ahat-vs-a type of analysis could not be performed, which had not been proposed at the time of the Bishop's study. Recording of the raw signals and the decision threshold could have

provides diagnostic capability for evaluations of potentially anomalous inspector calls, and it enables subsequent analyses not currently envisioned in the current study (e.g., using different thresholding techniques). As an example, for eddy current (a signal-response method relying on a threshold value) it is recommended that the peak voltage signal of an indication is recorded with the inspector's determination of whether to call a hit or miss that may include a subjective analysis of the Lissajous shape of the waveform. For an image-based hit/miss method (e.g., radiographic) it is suggested that the inspection images are recorded, preferably in a digital format that includes the film/detector specifications, the source strength, and exposure time with the inspector's final determination of whether to call a hit or miss. While this discussion on data recording may seem obvious, experience has proven this to be a challenging and underappreciated aspect of conducting a POD study that often leads to regrets regarding what was recorded.

It is recommended that the indicated flaw location on the specimen is recorded for each inspector call to enable its correlation with the true state of whether the flaw is present or not. When a flaw is present, a hit is a true positive, and a miss is a false negative. When no flaw is present, a hit is a false positive, and a miss is true negative. Unless the indicated location on the specimen is recorded for each inspector call, the inspector calls cannot be properly classified and included in the appropriate POD or POF analysis.

A primary motivation of this guidebook is to provide an approach to estimate Standard NDE flaw sizes for inspection methods that are in common practice, but are not included in NASA-STD-5009B. For newer NDE methods that are moving from a development stage to common practice, thereby warranting a Standard NDE study, it is recommended that there is additional emphasis placed on a establishing a defined protocol and inspector training to ensure a uniform implementation that is representative of operational field inspections.

The relative merits of pursuing Standard over Special NDE should be considered at this stage. Special NDE can be thought of as a unique case of Standard NDE, where every inspector's detection capability is evaluated and must meet or exceed a specified detectable flaw size rather than a sample from the population of possible field inspectors. Depending on a method's maturity, Special NDE may be deemed more appropriate.

## **3.2 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.<sup>5</sup> This is in contrast to MIL-HDBK-1823A (2009), Section 4.5.2.a that suggests a philosophy of using multiple specimen types to cover a family of structural types inspected in operational field inspections. 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, as was the case with Bishop (1973) that used

---

extended the value of Bishop's dataset with more informative subsequent analyses. Furthermore, it is expected that many of the larger flaws inspected with eddy current would have saturated the signal and/or exceeded the length of the eddy current probe, which can be accommodated in current analysis techniques with censored data. Since the flaw specimens were destructively characterized, it is not possible to conduct additional inspections to obtain signal responses.

<sup>5</sup> Bishop (1973) used flat aluminum panels that were considered representative of the SSP orbiter components that would undergo field inspections.

aluminum specimens and serves as the source for the penetrant and radiography open surface, partly through crack sizes, see NASA/TM-20220013820 (Parker, et al. (2022)). 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. As an example of a conservative flaw type, fatigue cracks have traditionally been considered to be worst-case, conservative flaws in evaluating the POD of methods for surface imperfections in metallic components.

MIL-HDBK-1823A (2009), Section 4.5.2, Appendix F.2.2, and Appendix F.2.3 offer helpful guidance on test specimen considerations, fabrication, and flaw production, respectively. For NASA Standard NDE, it is recommended that naturally occurring or simulated induced flaws (e.g., fatigue cracks) that provide representative flaw-to-flaw variability are used instead of simulated fabricated flaws (e.g., electro-discharge machining (EDM) notches) that are expected to exhibit less flaw-to-flaw variability. While it would be ideal to utilize naturally occurring flaws, it is assumed that there will be an insufficient number of flaws that can be independently characterized and are available for a Standard NDE study. Therefore, induced flaws will be used in most Standard NDE studies, and they should be 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. Furthermore, it is recommended, where possible, that the method to induce the flaws mimics the fabrication and/or operational usage that may produce naturally occurring flaws. While open surface fatigue cracks are 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.

It is recognized that producing fatigue cracks of a nominal size is challenging, especially if constrained by multiple flaw parameters (e.g., crack depth at varying aspect ratios<sup>6</sup>). As noted in the NASA/TM-20220013820 (Parker, et al. (2022)), certain methods of flaw production can result in more open cracks (i.e., a wider crack opening that could influence detectability). The flaw production technique should be given careful consideration in terms of its impact on the NDE method's detection capability. Furthermore, it should not be solely relegated to the facility that produces the cracks, which may not appreciate the impact of the production method on detection capability. Detailed technical documentation of how the flaws are produced is recommended.<sup>7</sup>

### **3.3 Statistical Flaw Size Design**

Consistent with the philosophy of statistical design of experiments, the study is planned to meet specified objectives that are linked to the intended usage of the resultant Standard NDE flaw sizes. Based on the NDE method, the primary flaw size characteristic (e.g., depth) that will be related to POD is specified, and potential secondary flaw characteristics (e.g., aspect ratio) that may

---

<sup>6</sup> NASA/TM-20220013820 (Parker, et al. (2022)) discusses the various methods to produce the fatigue cracks in Bishop (1973) including bending, tension-tension, and sequential application of these methods to produce flaws of various lengths and depths at different aspect ratios.

<sup>7</sup> The failure to record and archive the load levels and cycles used to induce the flaws was cited as a significant regret by the participants in Bishop (1973) that were interviewed for the NASA/TM-20220013820 (Parker, et al. (2022)).

influence POD are identified.<sup>8</sup> Other statistical design parameters of the flaw 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. These design decisions are vital in supporting the POD modeling of individual inspector a90/95 flaw sizes. In this section, rationale and guidance on each of these design decisions are discussed.

The choice of the appropriate flaw characteristics (e.g., depth) related to POD is the first step in developing a flaw size design. As an example, for eddy current, flaw depth may be considered the primary characteristic that influences POD, and a maximum depth when saturation of the signal occurs may be specified. In some situations, a surrogate for the primary characteristic may be considered due to practical limitations. As an example, for dye penetrant, flaw volume that retains the penetrant fluid and flaw opening width may be considered the most influential characteristics that enhance an inspector's ability to visually detect a bright linear indication. However, flaw opening width and volume may be difficult to control in fabrication and to independently characterize. In this case, flaw face area based on an assumed elliptical (i.e., thumbnail) flaw shape may be used as a surrogate (e.g., Bishop (1973)), or a flaw length restricted to a minimum depth may be used.

In addition to the primary flaw characteristic, secondary flaw characteristics may be an important consideration in the flaw design to support POD modeling conditioned on other flaw parameters<sup>9</sup>. The NASA/TM-20220013820 (Parker, et al. (2022)) highlights the importance of considering the correlation among multiple flaw parameters that might influence the detection capability of a method in the planning phase.<sup>10</sup> If POD is dependent on multiple flaw parameters (e.g., detectable flaw depth depends on aspect ratio), then the flaw design needs to vary depth independently from aspect ratio to make the effects independently estimable. Otherwise, if depth is correlated with aspect ratio, then the two effects cannot be distinguished uniquely.

In the design of a Standard NDE POD study, it is expected that the flaw size design will be informed by previous POD studies, since a Standard NDE study would not typically be performed with an NDE method under development. The detectable flaw size for an individual inspector (i.e., a90/95) is expected to be estimated before embarking on a Standard NDE study. This prior knowledge of detection capability is a significant benefit in planning a Standard NDE study, and enables a strategic and efficient flaw size design. In some cases, previous studies may provide individual inspector detection capabilities from a small number of inspectors or only highly experienced inspectors. While these prior estimates are useful in planning the range of flaws sizes, an unbiased representative sample of inspectors may show increased inspector-to-inspector variability. In general, it is recommended that prior knowledge from previous POD studies should be leveraged to the greatest extent possible.

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

---

<sup>8</sup> NASA/TM-20220013820 (Parker, et al. (2022)) showed that flaw size design choices resulted in correlation between flaw characteristics (e.g., flaw area and aspect ratio) that restricted the modeling of POD.

<sup>9</sup> As an example, NASA-STD-5009B's Standard NDE flaw sizes for penetrant suggest that the detectable area depends on aspect ratio.

<sup>10</sup> NASA/TM-20220013820 (Parker, et al. (2022)) showed that the Bishop (1973) flaw size design does not support the estimation of the effect of aspect ratio on penetrant's detectability, since there are too few large flaws at the low aspect ratio and too few small flaws at the higher aspect ratio. Even though Bishop's study included flaws with different aspect ratios, the distribution of flaw sizes with respect to aspect ratio was insufficient to estimate the effect.

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. (2012) Section 6.2 discusses the results of an extensive simulation study and recommends a maximum flaw size at the a97 (POD of 97%), and a minimum flaw size at the a3 (POD of 3%), which are flaw sizes that may be estimated from prior 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. This is an important distinction when considering MIL-HDBK-1823A's explicit focus on estimating the entire POD curve. MIL-HDBK-1823A (2009), Section 4.5.2.2.a recommends uniformly spaced flaw sizes between the maximum and minimum flaw sizes in the study. If a transformation of the flaw size is anticipated (e.g.,  $\log(\text{size})$ ), then Annis et al. (2012) Section 6.2 recommends that the distribution of flaw sizes are uniform based on the transformed flaw size. This results in a right-skewed distribution of flaw sizes in their original units (i.e., before transformation). A right-skewed distribution features a concentration of smaller flaws that are effectively stretched more significantly than larger flaws when applying a logarithm transformation, tending toward a uniform distribution of flaw sizes in the transformed flaw size.

MIL-HDBK-1823A suggests that a concentration of flaws around the a90 region may be beneficial, and it suggests a concentration of flaws in the transition region near a50, which is supported by Safizadeh et al. (2004). Concentrating flaws around a90 is consistent with the NASA Standard NDE study primary objective of estimating the a90/95 flaw size rather than entire POD curve as used by some commercial and military operations. 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. (2022) defines an approach to quantify the overlap as the percentage of flaws between the smallest hit and largest miss, which may be estimated from prior POD studies. 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 et al. (2022) 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 (e.g., for penetrant) once a flaw becomes so shallow, even if it is very long to increase the retaining volume, there may be an insufficient reservoir to retain a sufficient amount of penetrant for the flaw to be readily detectable. 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 is 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 the study depends on whether the inspector reports a signal-response or a hit/miss call. MIL-HDBK-1823A (2009), 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.<sup>11</sup> 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 et al. (2022) 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 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. It is assumed that the 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.

It is recommended that replicated flaws of the same nominal size throughout the range of flaw sizes be included in the study. Replication provides a straightforward means to assess the NDE method's detection variability when presented with multiple flaws of the same nominal size. Including replication is a best practice in statistical experiment design, and for signal-response methods it enables pure-error estimates that are helpful in evaluating the sufficiency of the response versus flaw size regression model. NASA/TM-20220003648 (Koshti et al. (2022)) provides a discussion on replication and pure error in the context of NDE POD modeling. While MIL-HDBK-1823A recommends 40 and 60 flaws for signal-response and hit/miss methods, respectively, it does not make a recommendation regarding the number of unique flaw sizes and whether replicate flaws are recommended.<sup>12</sup> A minimum of 3 replicates of the nominal flaw size allows for a model-independent estimate of variability. As an example of how replication might be employed in the study design, for a signal-response method, 40 flaws might be allocated as 4 replicates of 10 nominal flaw sizes across the flaw size range.

Considering a hit/miss method, replication of nominal flaw sizes is conceptually similar to NASA-STD-5009B's point estimate method (PEM) flaw design based on Rummel (1982), where there are 29 flaws on the same nominal size. However, when considering replication of flaws over a range of sizes, fewer replicated flaws is sufficient. For a hit/miss method, the total number of 60 flaws might be allocated as 4 replicates of 15 unique flaw sizes distributed across the flaw size range. It is recognized that exact replicates are unlikely to be obtained for some flaw types (e.g., induced fatigue cracks). However, near-replicates are practically useful.

It is recommended that the set of flawed specimens be specifically designed for the NDE method under study.<sup>13</sup> In some cases, a single set of flaws could be considered for multiple NDE methods that have similar detection capability and the primary flaw characteristic related to POD. While attractive logistically, using the same set of specimens for multiple NDE methods may prove

---

<sup>11</sup> Bishop (1973) featured 420 fatigue cracks, which is excessive relative to MIL-HDBK-1823A's recommendations. However, this large number of flaws may have been driven by the binomial analysis methodology employed in the early 1970s coupled with the need to span the detectable flaw size across the 4 methods (i.e., radiographic, penetrant, ultrasonics, eddy current) with a single set of flaws.

<sup>12</sup> Bishop (1973) included as many as 14 replicates at some flaw depths used in the eddy current POD modeling.

<sup>13</sup> NASA/TM-20220013820 (Parker, et al. (2022)) suggests that the Bishop (1973) study flaw size design was developed to span the detectable flaw sizes of the 4 methods (i.e., radiographic, penetrant, ultrasonics, eddy current) in a single set of flawed specimens, some of which have dramatically different detectable a90/95 flaw sizes. As a result, for eddy current, there were too many large flaws that were not particularly beneficial for modeling the POD, and conversely, for radiography there were too many small flaws and not enough larger flaws to avoid extrapolation.

challenging to avoid compromising the quality of the POD modeling of the individual NDE methods.

Unflawed specimens/sites need to be included in the study to preclude inspector guessing and to estimate the POF.<sup>14</sup> 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-5009C is expected to require that POF for Special NDE be reported, but no maximum POF is specified.

The Limited Sample POD (LS-POD) methodology, NASA TM-20210018515 (Koshti et al. (2021)), suggests a 1% POF with 95% confidence. In terms of existing guidance for the number of unflawed specimens/sites, LS-POD recommends a minimum of 40 unflawed specimens/sites to estimate 1/95 POF. For a hit/miss NDE method, MIL-HDBK-1823A (2009), Section 4.5.2.2.c recommends including 3 times as many unflawed inspection sites as flawed sites, without supporting rationale. For a Standard NDE study with a minimum of 40 to 60 flaws, this would result in 120 to 180 unflawed specimens/sites. This unflawed specimen/site quantity appears to be a conservatively large number, and it may add a significant burden in execution and data recording. Traditionally, the acceptable POF level is based on the cost of incorrectly classifying parts and removing them from service or production. This is difficult to assess in general for Standard NDE flaw sizes that may be applied to many future applications. In the absence of an existing NASA POF requirement for Standard NDE that supports the statistical basis for the number of unflawed specimens/sites in a hit/miss method, the following rationale was developed.

The POF analysis approach in MIL-HDBK-1823A (2009), Appendix G.4.6 considers a binomial distribution that is conceptually the same as the PEM (i.e., 29/29). For a POD demonstration using PEM, a minimum of 29 flaws are inspected and no missed detections are allowed to demonstrate at least 90/95 detection capability. It follows that a 10% maximum POF estimate with 95% confidence requires 29 unflawed specimens/sites to be inspected with no false positives allowed. While this illustrates how the approach and sample size are similar to a PEM, 10% POF is considered to be high for Standard NDE. Furthermore, MIL-HDBK-1823A suggests a 50% confidence level for the estimated POF, rather than 95%. MIL-HDBK-1823A (2009), Table G-I illustrates the relationship between the number of unflawed specimens/sites, the number of false positive calls, and the confidence level in estimating POF. If there are no false calls in 60 unflawed sites, then a maximum of 1% POF with 50% confidence is demonstrated. To demonstrate about 1.5% POF with 95% confidence, 200 unflawed specimens/sites are required, which appears to be a conservatively large specimen number. Therefore, drawing on the 1% POF recommendation from LS-POD for Special NDE of a signal-response NDE method, and changing the POF confidence level to 50% based on MIL-HDBK-1823A, it is recommended that a minimum of 60 unflawed specimens/sites with no false positive detections be employed in a Standard POD study for a hit/miss method. In summary, it is recommended that a minimum of 40 unflawed specimens/sites are used for a signal-response method and a minimum of 60 unflawed specimens/sites for a hit/miss method.

The number of unflawed specimens may take the form of unflawed regions of a specimen that contains one or more flaws, and thereby could be referred to as unflawed sites. For example, a grid of inspection locations could be marked on a single specimen, where some grid locations contain

---

<sup>14</sup> A significant weakness cited in Bishop's POD study was the lack of unflawed inspections included in the design to evaluate the POF.

flaws and others do not. If this approach is taken, a flaw location should not be deduced for each specimen (e.g., the flaw should not always be near the center of the specimen nor should flaws be uniformly spaced apart). The size of an inspection grid area should be considered representative of the NDE method's detection capability (e.g., relative to the largest flaw size in the study or related to a method's spatial resolution) and/or limitations in field inspections. The grid area should not be reduced in an effort to artificially increase the number of inspection opportunities on a single specimen. It is recommended that grid locations are marked on the specimen or an associated diagram for the inspector to reference, rather than inferred based on the area of the specimen that is inspected. If there are multiple inspection zones on a single specimen, then the inspector must report the zone where an inspection call is made.

The statistical flaw design decisions outlined in this section are expected to involve trade-offs of multiple qualitative and quantitative criteria, and will need to consider the estimated resource demands. Counter to conventional thought, a Standard NDE study does not need to be large if it is well designed.<sup>15</sup> The statistical design of the flawed specimens supports estimates of the a90/95 flaw size for individual inspectors and, when combined with the inspector sampling plan discussed in the following section, supports the estimation of the Standard NDE flaw size.

### **3.4 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 specific inspector population of interest to participate in the Standard NDE study. If sampled appropriately from the specific population, then the POD detection capability from this sample of inspectors can be used to infer 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 Level 2 or higher certification of inspectors in which Standard NDE flaw sizes are assumed detectable. Therefore, NAS-410 certified inspectors might represent the specific population of interest. Within the population of NAS-410 inspectors there may be an additional constraint to those inspectors that more routinely inspect aerospace components. Salkowski (1995) 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). As another case, the population of inspectors may be defined by the routine inspection of specific aerospace components. For example, 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. The population may be defined by an organization (e.g., inspectors within a specific facility or contractor) or by a collection of facilities

---

<sup>15</sup> Based on Bishop's comprehensive study, there was an implicit assumption that a Standard NDE study requires a larger number of flaws than a traditional POD study. However, a more modest number of flaws may allow Standard NDE studies to be more commonly performed.

and contractors supporting a single contract or program, as was the case in the Bishop (1973) study involving the SSP prime contractors.

Current NASA Standard NDE flaw sizes are applicable to inspections of fracture-critical spaceflight hardware and qualified inspectors inspecting this hardware, rather than inspectors associated with a specific program. In the near term, it is expected that new or updated Standard NDE flaw sizes will be applicable to the current population of qualified inspectors, and the inspector sampling may need to address this broad population of candidate inspectors. In the future, with Fracture Control Board approval, there may be a limited or restricted Standard NDE approach used for a specific application (e.g., COPV liners) or a specific company/facility. In some cases, there may be secondary features considered to define the population of inspectors that involves the NDE method implementation.<sup>16</sup> The defined inspector population may restrict or enable the application of the resulting Standard NDE flaw sizes in specific applications.

Once an inspector population is defined, a strategy for selecting a sample from that population is required. For the sample to be representative and unbiased, random sampling of inspectors and facilities is recommended, as described in Appendix C. Random sampling reduces bias and helps to ensure that there is a variety on inspector experience and capability included in the study. It might be enticing to select only the ‘best’ inspectors for a Standard NDE study. However, the resultant Standard NDE flaws sizes inferred from these inspectors would not be representative of the candidate inspectors that might perform operational field inspections on flight hardware. For longevity of the Standard NDE flaw size usage, 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.

After a population and sampling strategy are chosen, determining the number of inspectors to include in the Standard NDE POD study is required. Recall that estimating inspector-to-inspector variation is a unique objective of a Standard NDE, and therefore it follows that including more inspectors will provide more information of the detectable flaw size variability and increase the precision (i.e., reduce the uncertainty) 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 (1973), between 5 and 7 inspectors per NDE method were chosen to inspect 420 flaws. In this example, there are numerous 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. (1978) featured numerous inspectors and a relatively small number of flaws. Koh and Meeker (2017) explored a subset of this USAF database with 98 inspectors performing eddy current inspections on 52 flaws. Comparing Bishop (1973) to Lewis et al. (1978) 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,” (Brausch et al. (2008)) recommends “...at least 10-percent of the inspector population or at least 10 inspectors be included in the experimental design, whichever is larger.” As discussed, it is expected that in military

---

<sup>16</sup> In Bishop (1973) inspectors were chosen from 3 facilities, some of which employed automated eddy current scanning and recording techniques, and others of which used manual scanning and recording.

operations there will likely be many facilities and numerous inspectors at each facility, and it is expected that for 10 will be the larger value for NASA applications.

Based on the rationale presented in Appendix C, it is recommended that a minimum of 10 inspectors are chosen from the representative population of interest to conduct a Standard NDE study.<sup>17</sup> This minimum number of inspectors is considered reasonable based on NASA programs, and is based on the statistically consistent rationale contained in NASA/TM-20210018515 (Koshti et al. (2021)) and Spencer (2020a). It is recognized that there will be cases when the minimum number of inspectors cannot be obtained, and the analysis approach provided in the guidebook can be utilized when consulting a statistician and Fracture Control Board review. With a small number of inspectors, a Special NDE approach may be deemed more appropriate.

As with other study design decisions, the procedures and rationale for the inspector sampling plan should be documented. In particular, the definition of the representative population directly influences the appropriate application of the resultant Standard NDE flaw sizes. A primary objective of a Standard NDE study is to estimate inspector-to-inspector variability, and it is desirable to feature a large sample of inspectors to characterize the population of inspectors that could perform flight hardware inspections. 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.

### **3.5 Independent Flaw Size Characterization**

It is recommended that a strategy for the independent characterization of the flaw sizes is developed in the early planning stage of the POD study. Traditional 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. Notionally, for a hit/miss method with a binary response, random uncertainty in the flaw size will tend to flatten the slope of the transition region, which is a function of variability, in much the same way that inspector variability affects it. In this simple example, flattening of the transition region would increase the estimated a90/95 flaw size. While this simple example might imply conservatism, the reader is cautioned from taking this example as a general result, because the difference between the nominal size and the flaw size may include bias and random uncertainty where the effects are not readily predictable. While the recommendation to plan for independent flaw characterization may appear intuitive, experience has shown that it can be overlooked, and its value underestimated until after the flaw specimen production is completed leaving insufficient resources remaining to perform.

As discussed, a fatigue crack is anticipated to be the most common flaw type used in a NASA Standard NDE study. Independently characterizing the crack's length, depth, opening width, and shape (e.g., elliptical, thumbnail) is desired. Some flaw characteristics are more critical than others depending on the NDE method being studied. Historically, independent characterization was performed destructively by breaking open the flaw specimens and measuring their dimensions. In

---

<sup>17</sup> In Bishop (1973), there were 5 eddy current and ultrasonics inspectors, and 7 penetrant and radiographic inspectors, which fell below the recommended minimum number of 10. With only 5 inspectors, there is a significant small sample penalty, which may or may not be conservative, incurred in estimating the Standard NDE flaw size.

some cases, a sample of the same nominal flaw size are destructively tested to infer the size of a group of flaws that are nominally the same size. For a Standard NDE study, it is recommended that every flaw be independently characterized.<sup>18</sup> Making general inferences about the difference between the sampling of intended versus measured flaw sizes across a dispersed flaw size range is inherently more difficult to defend.

If destructive flaw characterization is used, then it should occur after a preliminary assessment of the inspection data quality. As part of this pre-characterization analysis, it might be valuable to perform POD modeling based on partial measurements that can be obtained nondestructively (e.g., using a scanning electron beam microscope) of the crack surface features (e.g., flaw length). Furthermore, it is recognized that some characteristics of the flaw's as-inspected condition may not be able to be measured destructively (e.g., crack opening width<sup>19</sup>). Every opportunity should be afforded to investigate anomalous inspection results before the specimen is destructively examined.

With advancements in computed topography (CT), independent characterization of some flaws may be performed nondestructively. There are clear advantages of measuring the flaw size in its as-inspected state, and it allows for the specimens to be re-inspected in future POD studies, which maximizes the value of the specimen production investment. However, the flaws may need to undergo some level of proof-testing to aid in a CT characterization. This loading plastically deforms the crack making it more 'open' to aid in CT measurement. Note if the crack is deformed in this way, then the crack opening width is no longer in the as-inspected condition, and it would preclude its' future usage in subsequent POD studies.

Independent flaw characterization is critical to POD modeling, and it may represent a significant proportion of a Standard NDE study's resource demands. Clearly, it is influenced by the number of flaws included in the study, and this provides additional motivation to specify a sufficient, but not excessive, number of flaws.

## **4.0 Execution**

It is recommended that the execution protocol of a Standard NDE study be documented and independently monitored. MIL-HDBK-1823A (2009), Section 4.5.3 recommends that a test monitor is designated to assure that guidance provided in the execution protocol is followed. It is recommended that a designated test monitor for the Standard NDE study is present at each facility during the inspection process. This independent oversight of the inspection process improves the reproducibility and validity of the resultant Standard NDE flaw sizes.

It is recommended that a briefing be developed to provide consistent instructions to the inspectors and/or facilities participating in the study. Without adherence to the instruction provided, one group of inspectors may inadvertently gain an advantage in the inspection process. In the USAF's "Have Cracks, Will Travel" program, Lewis et al. (1976) 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. First, a management briefing provided a description of the program goals and a discussion of the engineering and scientific

---

<sup>18</sup> Bishop (1973) independently characterized every flaw with destructive testing.

<sup>19</sup> Bishop (1973) assumed that flaw area could be used as a surrogate for flaw volume.

technologies. Second, an inspector briefing provided specific instructions on how to conduct the inspections and record their findings. It is recommended that a “dry run” with a limited number of inspectors is performed to ensure the briefing adequacy, and eliminate confusing elements and/or identify omissions.

Before executing the Standard NDE study, it is recommended that 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, it is recommended that a primary set of specimens be 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.

Since a Standard NDE study will likely involve shipping of specimens between different facilities, it is recommended that custom-designed shipping containers are utilized where each specimen has a designated location and is protected from mechanical damage or contamination during shipment. Intermittent physical inspections of specimens should be performed to detect changes that might affect an inspector or the method detection capability. At a minimum, pre- and post-shipment inspection and documentation are recommended to be compared to the baseline physical inspection to ensure that no damage or wear has occurred.

A cleaning process should be implemented after every inspection, and the materials and protocol should be carefully considered to not damage the specimen or influence an inspector or method’s detection capability. Based on the recommended minimum number of 10 inspectors, every specimen will be inspected and cleaned at least 10 times. Maintaining the integrity and original condition of the specimens is a strong assumption in comparing the first inspector’s detection capability to the last inspector’s detection capability. If specimen/flaw degradation occurs over time, then this should be noted in the periodic inspections and those specimens may be treated differently in POD analysis. The documentation of the inspectors’ order correlated with the specimen inspections and cleaning timeline is vital to consider this information in the analysis. MIL-HDBK-1823A (2009), Section 4.5.2.3 and Appendix F.2.3 offer helpful guidance on specimen maintenance.

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. It is recommended that 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.

It is recommended that noninformative specimen designations are assigned randomly. 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 designation, then 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. Association of the unique specimen identifier with the specimen characteristics (e.g., flaw size) should be available only to the test monitor or proctor, NDE engineer, and/or statistician overseeing the study.

Consistent with MIL-HDBK-1823A (2009), Section 4.5.2.c.1, it is recommended that a training set of specimens be designated. These training specimens allow an inspector to practice and

optimize their inspection technique, and become familiar with the specimen geometry before conducting inspections on the primary set of specimens used in the study. These practice inspections do not need to be performed in a blind manner. The main purpose of the training specimens is to minimize a learning-curve effect that otherwise might benefit the inspections later in the execution order compared to the first specimens presented the inspector.

As a supplemental consideration, there are benefits to presenting an inspector with some of the same flawed and unflawed specimens/sites multiple times to evaluate within-inspector repeatability, if this can be accomplished without the inspector's awareness that they are the same specimens. These inspections are known as repeated measurements, which is different from the recommendation in Section 3.3 to include multiple specimens with the same nominal size flaw, known as replicated flaws. Conceptually, when presented the same specimen with low POD flaws or unflawed specimens/sites, the inspector should report no indication consistently. Similarly, for extremely large flaws, an inspector should consistently find the flaw. However, for flaws in the transition region, near a50, an individual inspector may only find the same flaw, if present, 50% of the time. Repeated presentations of the same flaws near the a50 size and larger toward the a90 size provide the most useful information. While the use of replicated flaws (i.e., multiple specimens with the same nominal flaw size) can provide similar information on inspector variability, repeat measurements on the same flaw removes the effect of flaw-to-flaw variability. Performing repeat inspections of the same specimen unbeknownst to an inspector may be challenging or infeasible. However, if an automated scanning or detection system is used in the POD study, repeated inspections can be accommodated. For some NDE methods (e.g., penetrant) the specimen would need to be cleaned during the inspection process of a single inspector before being presented to the inspector for a repeat inspection and this may preclude the practicability of repeated measurements. Finally, repeated inspections do not reduce the recommended number of flaw specimens in Section 3.3 since they do not provide independent information on flaw-to-flaw variability.

It is recommended that inspector fatigue due to conducting too many sequential inspections is considered in the execution protocol. This recommendation applies to manual methods of inspection and the 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, if there is an attempt to mimic the physical posture of the inspector during field inspections, then this may contribute to the inspection duration consideration. For a Standard NDE study, the primary objective is to achieve representativeness rather than seek improvement of detection capability by varying the operator's inspection interval and duration as would be done in a human factors study. While a study to determine the best inspection period is valuable, it would be performed external to a Standard NDE study to establish the protocol that will be implemented in the operational field inspections.

## **5.0 Analysis**

The analysis of NASA Standard NDE study is a two-step process that begins with the industry standard practice contained in MIL-HDBK-1823A for analyzing individual inspector detection

capability and is followed by analysis of individual inspector a90/95 flaw sizes to estimate a Standard NDE flaw size. 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 (1973) used a conceptually similar two-step approach, so it is historically consistent.<sup>20</sup> Alternative statistical approaches were considered, which are discussed in Appendix D.

As stated, the Standard NDE flaw size represents the a90/95 flaw size detectable by most inspectors in the population of interest. 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 that are expected to demonstrate 90/95 detection of the Standard NDE flaw size 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.

Historically, Bishop (1973) defined Standard NDE as 95% of the inspectors within the population of interest are considered to possess at least 90/95 detection capability of the Standard NDE flaw size.<sup>21</sup> However, NASA/TM-20220013820 (Parker, et al. (2022)) demonstrated that Bishop's estimates were generally nonconservative and did not represent 95% inspector coverage consistently. Spencer (2020b) suggests 90% inspector coverage as being analogous to the required 90% probability of detection, since both are a proportion of a population: one of flaws and the other of inspectors. For a given individual inspector POD model and a fixed number of inspectors, increasing the inspector coverage proportion increases the Standard NDE flaw size. As a minimum, 80% is considered a reasonable lower value for the proportion of inspectors, beyond which the fraction of inspectors that would not possess at least 90/95 detection of the Standard NDE flaw size seems too high for fracture analysis. In light of these considerations, it is recommended that 90% inspector coverage at 50% confidence based on individual inspector a90/95 flaw sizes is reasonable in the absence of specific requirements. The statistical analysis presented herein can be adapted for other levels of inspector coverage and confidence levels.

It is appropriate that a significant portion of this guidebook has been devoted to the planning of a Standard NDE study because careful attention to the study design and execution protocol are critical to support insightful and statistically defensible analyses. While not always explicitly stated, the design aspects discussed are tailored to support the recommended analysis method in this section. Regrettably, it is a common misconception that the study design and analysis aspects are independent, which often leads to analysis limitations after the specimens are fabricated and the inspections are completed. As shown in the NASA/TM-20220013820 (Parker, et al. (2022)), analysis techniques may not be able to overcome deficiencies in the design. In summary, the design

---

<sup>20</sup> Bishop used the sorted group ascent method (SGAM) to estimate individual inspector capability, rather than POD modeling approaches contained in MIL-HDBK-1823A. See NASA/TM-20220013820 (Parker, et al. (2022)) for SGAM information.

<sup>21</sup> NASA/TM-20220013820 (Parker, et al. (2022)) showed through reanalysis of the Bishop POD dataset that limited cases met or approached 95% inspector coverage probability. In many cases, the coverage probability was 50% or less, which indicates that as many as one-half of certified inspectors would not be expected to demonstrate a 90/95 detection capability of the Standard NDE flaw size.

and analysis are highly interdependent, and a holistic approach to Standard NDE considers the data required to enable the intended analyses.

## **5.1 Estimating Individual Inspector a90/95**

A Standard NDE study can be thought of as a collection of individual inspector POD studies. If every inspector sees every specimen, then the study is a fully crossed design. 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 the most common in practice. In a fully crossed design, the inspection results from an individual inspector can be analyzed using MIL-HDBK-1823A methods to estimate an individual inspector's a90/95 flaw size. This handbook addresses signal-response and hit/miss NDE methods, and provides guidance on the aspects of the analysis that include data quality diagnostics, the determination of flaw size and/or signal transformation, choice of an appropriate distributional model, choice of the link function (e.g., logit or probit) for the generalized linear model regression, and an analysis of noise to estimate the probability of false calls. MIL-HDBK-1823A's guidance is considered sufficiently detailed to estimate an individual inspector's a90/95 flaw sizes in most cases, and is not described here. While not addressed in MIL-HDBK-1823A, it is recommended that a consistent modeling approach be used for the inspectors in a Standard NDE POD study, rather than individual models for each inspector. Therefore, the best modeling approach across the inspectors, including flaw size transformations, is recommended.

While the basic 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. For example, in some cases the minimum probability of detection may not be zero due to background noise of the NDE method. In addition, the maximum probability of detection may not be equal to 1, especially if there are misses unrelated to flaw sizes. In this case, the POD model may feature an estimated upper and/or lower asymptote, as discussed by Spencer (2014) and MIL-HDBK-1823A (2009), Appendix I.4. Furthermore, POD models that are a function of multiple flaw characteristics may be appropriate as discussed from a design perspective in MIL-HDBK-1823A (2009), Appendix E. For example, the POD of a method may be related to the flaw length and depth, or it may be related to flaw area and aspect ratio. As discussed in Section 5.0, considerations are required in the statistical flaw design to support the estimation of these multi-factor models. If the POD relationship depends on multiple flaw characteristics, then the Standard NDE flaw size may be reported as contingent on a secondary flaw characteristic (e.g., the detectable flaw length for a specific aspect ratio). These more complex POD models are considered beyond the scope of this guidebook, and consultation with a statistician is recommended.

## **5.2 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. In Section 3.3, it was suggested that a maximum of a 1% probability of a false positive at 50% confidence (i.e., 1/50) should be demonstrated by each inspector in the absence of a specific requirement. False positives can occur in signal-response and hit/miss NDE methods, and their distinctive analysis approaches for each method type are subsequently described.

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 is employed to estimate the upper bound on the signal associated with 99<sup>th</sup> percentile (i.e., 1% upper quantile) with 50% confidence. For every individual inspector, it is recommended this signal level be reported. 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. This is similar to the approach described in the LS-POD Special NDE method, see NASA/TM-20210018515 (Koshti et al. (2021)). However, a more conservative 1/95 POF is utilized in a LS-POD demonstration. Computing the 1/50 signal level and comparing it to the decision threshold is the most straightforward approach to ensure the inspectors demonstrate a maximum of 1/50 POF. Alternatively, the average and standard deviation of the unflawed signals can be used to estimate the POF with a 50% confidence based on the signal decision threshold. In this approach, the estimated POF should not exceed 1%.

For a hit/miss method, the number of false positive indications is recorded and compared to the total number of unflawed specimens/sites inspected. MIL-HDBK-1823A (2009), Appendix G.4.6 describes this analysis for estimating POF. If 60 unflawed specimens/sites are included in the study, then no false positives are allowed to demonstrate 1/50 POF.

For signal-response and hit/miss NDE methods, in the absence of a specific requirement it is recommended the individual inspectors in a Standard NDE study demonstrate a maximum of 1/50 POF. If an inspector exceeds the recommended 1/50 POF, a diagnostic phase is suggested to identify a correctable cause. If a correctable cause is not found, an inspector that exceeds 1/50 POF may be included in a Standard POD study under the review of the Fracture Control Board.

### 5.3 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 ( $\overline{a_{90/95}}$ ) and the standard deviation of the a90/95 flaw sizes across the inspectors ( $s_{\text{insp}}$ ) 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.

A relatively small number of inspectors (i.e.,  $\geq 10$ ) is expected to be included in a NASA Standard NDE study. Therefore, it is most likely impractical to reliably identify the appropriate underlying distributional model of individual inspector a90/95 flaw sizes and a Normal distribution is assumed. If the sample size is large enough, then the appropriateness of a Normal distribution could be evaluated, and other distributions considered in the tolerance interval computation. It is recommended that the distribution of individual inspector a90/95 flaw size estimates is examined to identify anomalously small or large detectable flaw sizes and/or distinct clusters of flaw sizes that may warrant further investigation before including them in the calculation of a Standard NDE flaw size.

Since the estimated Standard NDE flaw size is based on a sample of inspectors from the population of interest, a factor,  $k_1$ , compensates for the uncertainty associated with inferring the population characteristics. The smaller the number of inspectors used to infer the population characteristics, the larger the  $k_1$  factor.

Meeker et al. (2017), Equation 4.2 provides the formulation for a confidence interval on a normal distribution quantile as a function of the inverse cumulative distribution function of a normal

distribution and a non-central  $t$ -distribution, and it is adapted for a one-sided tolerance interval to estimate  $k_1$  as:

$$\delta = z_p * \sqrt{n} \quad \text{Equation (1)}$$

$$k_{1_{0.90/50\%}} = \frac{t_{(\alpha, n-1, \delta)}}{\sqrt{n}} \quad \text{Equation (2)}$$

where  $z_p$  is the critical value from a standard normal distribution for  $p = 0.90$  based on a 90% proportion of the population of inspectors,  $\alpha = 0.50$  indicates a 50% confidence level,  $\delta$  is the non-centrality parameter, and  $n$  is the number of inspectors. While the confidence level could be chosen as 95%, a 50% confidence level is recommended to avoid compounding of conservatism with multiple confidence levels as the computation includes a 95% confidence level of the individual inspector's 90% probability of detection.

The statistical tolerance interval calculation is provided in most statistical software packages. However, the  $k_{1_{0.90/50\%}}$  values for 7 to 20 inspectors are included in Appendix A. Note this Appendix illustrates the diminishing benefits with more than 10 inspectors, which supports the rationale for recommending this as the minimum number of inspectors.

The Standard NDE flow size that provides 90% probability of detection at 95% confidence for 90% of the inspectors from the population of interest is estimated by:

$$a_{90/95_{Std}} = \overline{a_{90/95}} + k_{1_{0.90/50\%}} * S_{insp} \quad \text{Equation (3)}$$

As discussed, alternative proportions of inspectors that are expected to possess at least 90/95 detection capability of the Standard NDE flow size can be computed using a different  $p$  in the  $k_{1_{p/50\%}}$  calculation.

As a technical note, if a logarithm transformation is applied to the flow size in the individual inspector POD modeling, then the variability in a90/95 flow sizes across the inspectors is recommended to be estimated with an untransformed flow size (i.e., in the original flow size engineering units). In brief, the common usage of transformations in the POD modeling for individual inspectors are empirically chosen to improve POD modeling, and are not related to the physics of the NDE method. If a transformation is used in the POD modeling, then the individual a90/95 flow sizes in their original units are reported as the inspector's capability.<sup>22</sup>

This recommended Standard NDE analysis approach is expected to be straightforward to implement, and its simplicity enhances the insights regarding the range of detection capability of individual inspectors sampled from the population of interest. Appendix B provides numerical examples to illustrate the analysis approach for hit/miss and signal-response NDE methods.

## 6.0 Documentation

Documentation of Standard NDE studies is vital for traceability of the reported Standard NDE flow sizes, reproducibility of the study, and assessments of similarity and transferability to specific flight components. Thorough documentation is especially important considering the expected

---

<sup>22</sup> This recommendation is based on independent analyses of Bishop (1973) POD study. For ultrasonics, though  $\ln(\text{area})$  was used in individual inspector POD modeling, it was found that using a transformed a90/95 in computation of the Standard NDE flow size resulted in an overly conservative Standard NDE flow size due to wide variability among the ultrasonic inspectors that was amplified by using a transformation.

longevity of applying the resultant Standard NDE flaw sizes, as evidenced by the results of Bishop (1973) that remain in broad usage after 50 years.<sup>23</sup> As with other research publications, independent reproducibility of the Standard NDE flaw size is a strong motivation. This is especially important when considering the potential impact on space flight systems safety and mass efficiency.

This guidebook is intended to serve as a template or checklist for the type of documentation that includes the design choices in the planning stage and their rationale. Of particular importance is the definition of the population of inspectors in which the Standard NDE flaw sizes are intended to be valid. The documentation should provide the necessary details of the POD modeling for individual inspectors and the estimation of the Standard NDE flaw size that would allow those results to be numerically reproduced.

As discussed, it is anticipated that most Standard NDE studies will utilize simple specimen geometries, and therefore, an evaluation of transferring the Standard NDE flaw sizes to the flight hardware will by necessity be required, as stipulated in NASA-STD-5009B. The Standard NDE flaw sizes are a baseline to compare differences in materials, geometry, and possibly flaw type to operational flight inspections. NASA/TM-20220003648 (Koshti et al. (2022)) provides guidance on conducting an evaluation of similarity and developing transfer functions to utilize the results of a Standard NDE study.

In addition to the comprehensive inspection report and POD analysis, a condensed summary specification is recommended. As an example, the NDE Capabilities Databook (1997) utilized a consistent template to summarize each individual POD study. This template was adapted and augmented for a NASA Standard NDE study in Figure 6.0-1 to serve as an example condensed specification of the study design. The numerical examples in Appendix B illustrate a minimum documentation of the analysis. The raw inspection data should be provided to allow for independent reproduction of the results.

---

<sup>23</sup> NASA/TM-20220013820 (Parker, et al. (2022)) highlighted poor or incomplete documentation through the evolution of NASA Standard NDE. This finding serves as a strong motivation to emphasize the documentation in this guidebook.

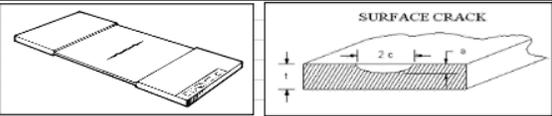
<b>NDE METHOD:</b>	Fluorescent Penetrant
<b>NDE PROCEDURE:</b>	Fluorescent Penetrant Manual - URESCO P149, Solvent Removable, Spray Developer
<b>NDE TECHNIQUE:</b>	Manual Inspection / Manual Recording
<b>NDE DATA TYPE:</b>	Hit / Miss, and Location of Call on specimen
<b>NDE METHOD THRESHOLD</b>	<i>(applies to signal-response methods including how the threshold was determined)</i>
<b>SPECIMEN MATERIAL:</b>	2219 Aluminum T-87
<b>SPECIMEN</b>	Flat Plate - 3.5 inches by 16 inches, cracks on both sides
<b>SPECIMEN THICKNESS:</b>	0.060 and 0.225 inch nominal
<b>SPECIMEN/FLAW CONDITION</b>	After Etch
<b>SPECIMEN SURFACE FINISH:</b>	125 and 32 RMS - representative of good machining practices
<b>FLAW TYPE:</b>	Fatigue Cracks, Open Surface, Partly Through
<b>FLAW PRODUCTION</b>	Shaped EDM starter notch initiation, grown in bending and tension / tension
<b>FLAW SHAPE</b>	Aspect Ratio ( $2c/a$ ) = 2 to 10, Depth-to-Thickness ( $a/t$ ) = 0.2 to 0.9
<b>FLAW SIZE VERIFICATION</b>	Destructive analysis and measurement
<b>NUMBER OF FLAWS</b>	90
<b>NO. OF UNFLAWED SPECIMENS</b>	60
<b>INSPECTOR POPULATION</b>	NAS-410, Level II Certified
<b>NUMBER OF INSPECTORS</b>	10, randomly selected across the facilities
<b>PERFORMING ORGANIZATION:</b>	Facilities X, Y, Z
<b>DATE:</b>	May 2022
<b>SPONSOR:</b>	Performed in support of NASA Program A, under Contract No. 1000
	

Figure 6.0-1. Example Standard NDE study condensed specification template.

## 7.0 Conclusions

The methodology proposed in this guidebook provides an approach to update the existing Standard NDE flaw sizes and add flaw sizes for new NDE methods. It leverages lessons learned from the seminal SSP studies, NDE literature, and extends MIL-HDBK-1823A's guidance to a NASA Standard NDE study. A unified approach to the statistical design and analysis of a Standard NDE POD study is presented to strategically meet the needs of envisioned applications of the Standard NDE flaw sizes through careful planning, execution, analysis, and documentation, and the guidance is summarized in a convenient checklist in Appendix E. The proposed method was developed to be straightforward, intuitive, and approachable to NDE practitioners and fracture analysts to broaden its potential application.

The methodology contained herein is the first documented NASA Standard NDE approach. While simulated numerical examples are provided to illustrate the approach, it is acknowledged that it has not been used to conduct a Standard NDE study. Once exercised in application, this initial methodology is expected to be refined and augmented. In summary, this guidebook addressed a significant, long-standing gap in the Standard NDE body of knowledge, and it supports the continued usage of Standard NDE flaw sizes in the majority of NASA's spaceflight system designs.

## 8.0 References

- 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.
- Bishop, C. R. (1973): "Nondestructive Evaluation of Fatigue Cracks," Space Division Rockwell International, SD 73-SH-0219.
- 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, October 2008.
- Burdick, R. K.; Borrer, C. M.; and Montgomery, D. C. (2005): "Design and Analysis of Gauge R&R Studies : Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models," American Statistical Association and the Society for Industrial and Applied Mathematics.
- 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.
- Koh, Y.M., and W.Q. Meeker (2017), "Quantile POD for Nondestructive Evaluation with Hit-Miss Data," *Research in Nondestructive Evaluation*, 30:2, October 2017, pg. 89-111.
- 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.
- 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.
- 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.
- Meeker, W. Q.; Hahn, G. J.; and Escobar, L. A. (2017): "Statistical Intervals: A Guide for Practitioners and Researchers," 2nd Edition, John Wiley & Sons Inc., 2017.
- MIL-HDBK-1823A (2009): "Nondestructive Evaluation System Reliability Assessment."
- NAS 410 (2020): "NAS Certification and Qualification of Nondestructive Test Personnel," 5th Edition.
- NASA Standard 5009B (2019): "Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components," 2019.
- NASA Standard 5019A (2019): "Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components," 2019.
- NDE Capabilities Data Book (1997), 3<sup>rd</sup> Edition, Nondestructive Testing and Information Analysis Center (NTIAC), prepared by Rummel, W. D., Matzkanin, G. A.

- 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.
- Rummel, W. D. (1982): "Recommended Practice for Demonstration of Nondestructive Evaluation (NDE) Reliability on Aircraft Production Parts," *Materials Evaluation*, 40.
- Safizadeh, M. S.; Forsyth, D. S.; and Fahr, A. (2004): "The Effect of Flaw Size Distribution on the Estimation of POD," *Insight*, 46, 6.
- Salkowski, C. (1995): "Nondestructive Inspection Reliability Assumptions for Critical Aerospace Components," *Proceedings SPIE 2455, Nondestructive Evaluation of Aging Aircraft, Airport, Aerospace Hardware, and Materials*.
- Spencer, F. W. (2014): "Spencer, F. W. (2014), "Curve Fitting for Probability of Detection Data: A 4-Parameter Generalization," 40th Annual Review of Progress in Quantitative Nondestructive Evaluation, AIP Conf. Proc. 1581, 2055-2062.
- Spencer, F. W. (2020a): "Recommendations on Inspector Numbers for a Standard Methodology of NDE Characterization," informal correspondence.
- Spencer, F. W. (2020b): "Response to questions/comments on Recommendations on Inspector Numbers for a Standard Methodology of NDE Characterization," informal correspondence.

**Appendix A – Tabulated  $k_1$  Values to Estimate Standard NDE Flaw Size**  
**Values for 90% inspector coverage at 50% confidence for 7 to 20 inspectors in the study.**

No. Insp	k1 90/50
7	1.347
8	1.337
9	1.330
10	1.324
11	1.320
12	1.316
13	1.313
14	1.311
15	1.309
16	1.307
17	1.305
18	1.304
19	1.302
20	1.301

## Appendix B – Standard NDE POD Study Analysis Examples

In this Appendix, simulated numerical examples are provided to illustrate Standard nondestructive evaluation (NDE) probability of detection (POD) study design aspects and analysis.

### Hit/Miss NDE Method Example

A Standard NDE POD study for a hit/miss NDE method is illustrated with 60 flawed specimens and 10 certified inspectors randomly sampled from a specific population of interest. Every inspector sees each of the 60 flaws once and calls either a hit (1) or miss (0) in a fully crossed design. The POD of this NDE method is assumed to be related to a single flaw size characteristic (e.g., flaw length). This example was inspired by MIL-HDBK-1823A (2009), Appendix G, Example 3 dataset.

The 10 inspectors are presented with 60 unflawed sites to estimate the probability of false calls (POF), and none of the inspectors reported any false calls. Therefore, they have demonstrated a maximum of 1% POF with 50% confidence, and the inspectors are included in the POD modeling and estimation of the Standard NDE flaw size.

Table B.1 shows the inspection results from the 10 inspectors for the flawed specimens. The flaw size units were independently characterized using destructive testing after the inspection process and preliminary data analysis were completed. The 60 flawed and 60 unflawed specimens/sites were randomly intermingled during the inspection process. The 60 no-flaw calls for the unflawed sites have been omitted from the data table for conciseness. The flaw specimens are assigned a unique, noninformative two-letter sequence unrelated to flaw size to avoid prompting the inspector to guess whether a flaw is present, nor guess its size if it is present.

Figure B.1 plots the calls from Inspector 1 versus the flaw size, as an example diagnostic plot, where each dot represents a single inspection. Replicated flaw sizes are indicated by stacks of dots. This plot illustrates the distribution of flaw sizes in the study. By inspection, there are small flaw sizes that are rarely detected and larger flaw sizes that are nearly always detected, and there is sufficient overlap in the transition region. Approximately 50% of the flaw sizes reside between the smallest hit and the largest miss, referred to as overlap, and this results in a desirable concentration of flaws in the a50 region. About 30% of the inspection results are misses, referred to as evenness, which is consistent with the recommendations in Section 3.3.

For each inspector, a POD model is estimated using a logistic regression model (i.e., logit link function) as a function of flaw size, denoted as  $x$ , without transformation of the flaw size, as:

$$\overline{POD} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x)}}$$

The estimated parameters of the model,  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , are used to predict the a90 and a90/95 flaw sizes for each inspector. Figure B.2 shows the 10 POD curves, one for each inspector, and the estimated a90 flaw sizes are indicated with circle markers. At each circle marker, there is an arrow extending to the right, and at the tip of the arrow is the a90/95 flaw size (i.e., the 90% probability of detection with 95% confidence). Figure B.3 provides an enlarged view of the POD models in the region of 90% POD to visualize the variability in the a90 and a90/95 flaw size estimates.

Table B.2 provides the model coefficients for each of the individual inspector POD models, and it tabulates the a90 and a90/95 flaw sizes shown in Figures B.2 and B.3. The average a90/95 is 0.212, and the standard deviation of the a90/95 values among the inspectors is 0.032. For 10 inspectors,

the  $k_1$  value for 90% inspector coverage with 50% confidence is found in Appendix A as 1.324. Using Equation 3 in Section 5.3, the Standard NDE flaw size is estimated to be 0.254, which is the a90/95/90 Standard NDE flaw size. In this example, the Standard NDE flaw size is smaller than the largest individual inspector a90/95 of 0.265, and it is approximately equal to the next largest a90/95 of 0.256. This outcome illustrates the 90% inspector coverage of the estimated Standard NDE flaw size.

Table B.1. Hit/Miss Standard NDE POD Example Dataset with 60 Flaws and 10 Inspectors

Index	Specimen ID	Flaw Size	Individual Inspector Calls									
			1	2	3	4	5	6	7	8	9	10
1	AG	0.117	1	0	1	1	1	0	1	0	1	0
2	AQ	0.423	1	1	1	1	1	1	1	1	1	1
3	GR	0.075	0	0	1	0	0	0	0	0	0	0
4	CD	0.271	1	1	1	1	1	1	1	1	1	1
5	DI	0.241	1	1	1	1	1	1	1	1	1	1
6	CW	0.273	1	1	1	1	1	1	1	1	1	1
7	AN	0.073	0	0	0	0	0	1	1	0	0	0
8	GA	0.167	1	1	1	1	1	0	1	0	1	1
9	CO	0.073	0	1	1	0	0	0	0	0	0	0
10	GL	0.110	1	0	0	1	0	1	1	1	1	0
11	BR	0.155	0	1	1	1	1	0	1	1	1	1
12	EI	0.205	1	1	1	1	1	1	1	1	1	0
13	FI	0.082	0	0	0	0	0	0	0	1	0	0
14	DY	0.085	0	0	0	0	0	0	0	0	0	0
15	GH	0.077	0	0	0	1	0	0	0	0	0	0
16	AT	0.076	1	1	0	0	0	0	0	0	0	0
17	EY	0.043	0	0	0	1	0	0	0	0	0	0
18	AD	0.221	1	1	1	1	1	1	1	1	1	1
19	DJ	0.172	1	1	0	1	1	0	1	1	1	1
20	GU	0.126	1	1	1	1	1	1	1	1	0	1
21	CN	0.137	0	1	0	1	1	1	1	1	1	1
22	ET	0.026	0	0	0	0	0	0	0	0	0	0
23	CF	0.222	1	1	1	1	1	1	1	1	1	1
24	DX	0.145	0	1	0	1	1	1	1	1	0	1
25	CI	0.131	0	0	1	0	1	1	1	1	1	0
26	AE	0.257	1	1	1	1	1	1	1	1	1	1
27	BV	0.140	1	1	1	1	0	0	1	1	0	1
28	FT	0.308	1	1	1	1	1	1	1	1	1	1
29	EW	0.060	0	0	0	0	0	0	0	0	0	0
30	CA	0.102	1	1	0	0	1	0	1	1	1	0
31	CZ	0.250	1	1	1	1	1	1	1	1	1	1
32	BT	0.063	0	1	0	0	0	1	0	0	0	0
33	BW	0.137	1	1	0	0	0	0	0	1	1	1
34	DW	0.345	1	1	1	1	1	1	1	1	1	1
35	FH	0.286	1	1	1	1	1	1	1	1	1	1
36	BP	0.030	0	0	0	0	0	0	0	0	0	0
37	FS	0.143	1	1	0	1	1	1	1	0	1	1
38	DM	0.269	1	1	1	1	1	1	1	1	1	1
39	BM	0.216	1	1	1	1	1	1	0	1	1	1
40	FD	0.147	1	0	0	1	1	1	1	1	1	0
41	BG	0.317	1	1	1	1	1	1	1	1	1	1
42	GE	0.142	1	1	0	1	0	1	1	1	1	1
43	DF	0.149	1	0	1	1	1	0	1	1	1	1
44	GC	0.077	1	1	0	0	0	0	0	0	0	0
45	BI	0.378	1	1	1	1	1	1	1	1	1	1
46	BB	0.240	1	1	1	1	1	1	1	1	1	1
47	EO	0.136	1	0	1	1	1	1	1	1	1	0
48	FG	0.147	1	1	1	1	0	1	1	1	1	1
49	EN	0.107	1	1	0	1	1	1	0	0	0	0
50	FC	0.074	0	1	0	0	0	0	0	0	0	1
51	DR	0.124	1	1	0	1	0	0	0	1	1	0
52	DH	0.213	1	1	1	1	1	1	1	1	1	1
53	AZ	0.145	1	1	1	1	1	0	1	1	1	1
54	EH	0.154	0	1	0	1	1	1	1	1	1	1
55	EJ	0.205	1	1	1	1	1	1	1	1	1	1
56	DK	0.132	1	1	1	0	0	1	1	1	1	0
57	CS	0.219	1	1	1	1	1	1	1	1	1	1
58	DV	0.192	1	1	1	1	1	1	1	1	1	1
59	AR	0.093	0	0	0	0	0	0	0	0	0	0
60	GZ	0.275	1	1	1	1	1	1	1	1	1	1

Note: Flawed specimen data only is shown. However, 60 unflawed sites were intermingled with the flawed specimens during the inspection resulting in no false positives from the inspectors.



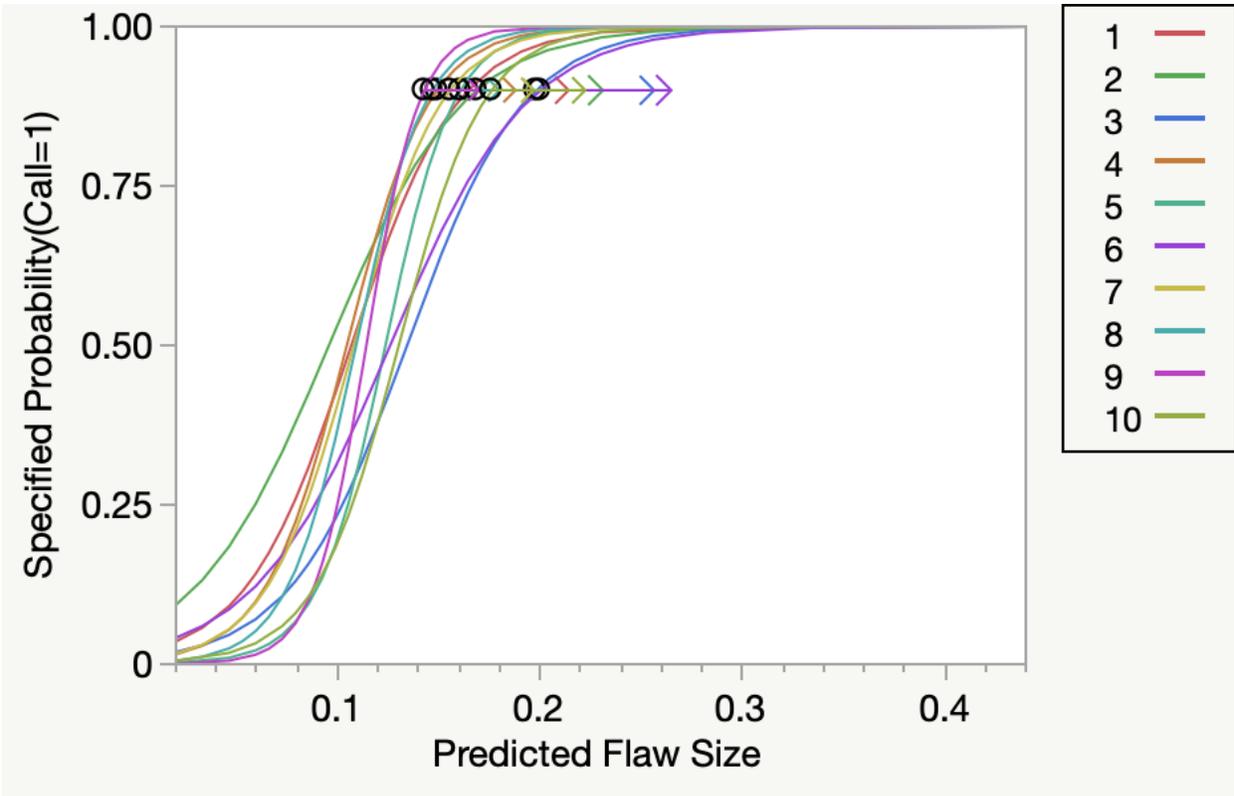


Figure B.2. Individual inspector POD models with a90 and 90/95 estimates for hit/miss standard NDE POD example.

Note: Circle markers indicate individual a90 flaw size and arrow tips indicate individual a90/95.

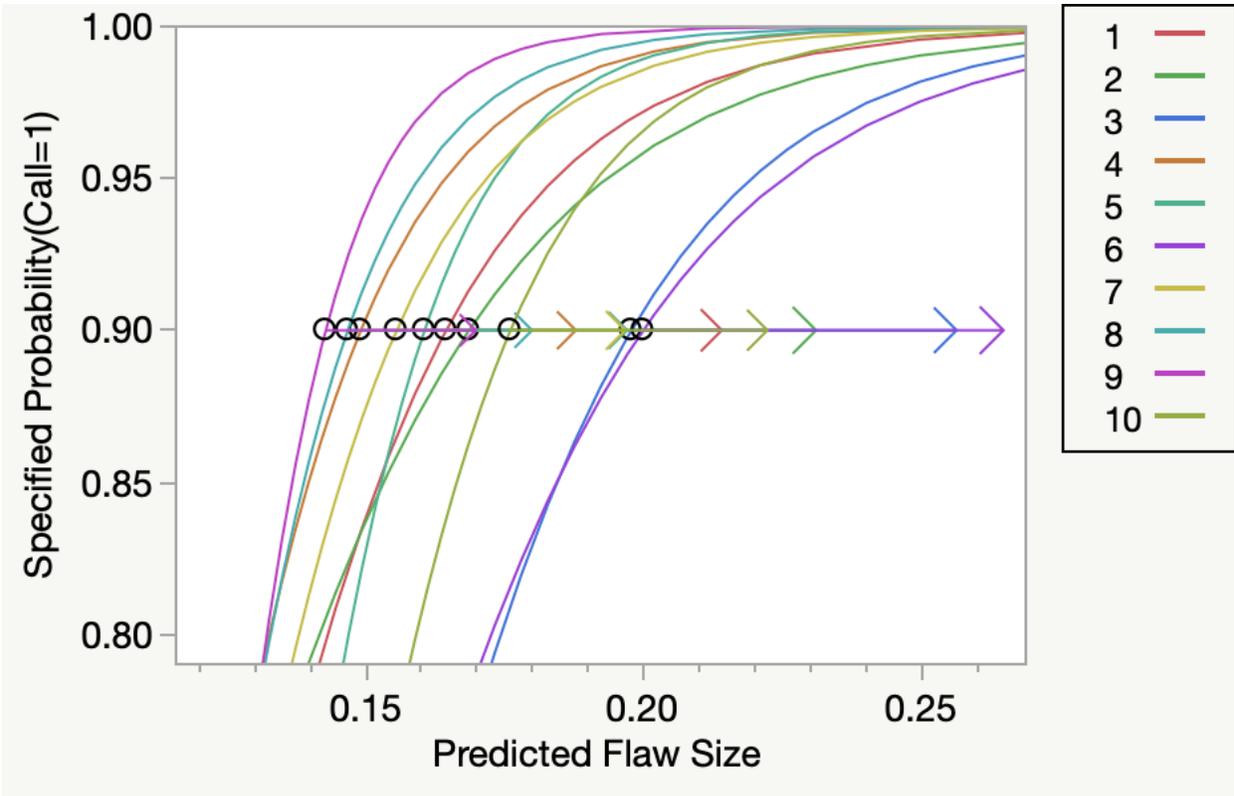


Figure B.3. 90% POD region of individual inspector POD models with a90 and 90/95 estimates for hit/miss standard NDE POD example.

Note: Circle markers indicate individual a90 flaw size and arrow tips indicate individual a90/95.

Table B.2. Individual Inspector Logistic Model Estimated Parameters,  $a_{90}$ ,  $a_{90/95}$ , and Standard NDE flaw size ( $a_{90/95/90}$ ) Flaw Size for Hit/Miss NDE Method Example

Inspector	$\beta_0\_hat$	$\beta_1\_hat$	$a_{90}$	$a_{90/95}$
1	-4.06	38.06	0.164	0.214
2	-2.88	30.12	0.169	0.231
3	-4.63	34.51	0.198	0.256
4	-5.14	49.20	0.149	0.188
5	-7.34	59.38	0.161	0.197
6	-3.73	29.67	0.200	0.265
7	-4.99	46.20	0.156	0.197
8	-6.37	58.40	0.147	0.180
9	-8.74	76.64	0.143	0.170
10	-6.23	47.85	0.176	0.222
			Inspector Average $a_{90/95}$	0.212
			Inspector-to-inspector variability, Stdev of $a_{90/95}$	0.032
			k1 value for 0.90/50% for 10 inspectors	1.324
			<b>Standard NDE Flaw Size, <math>a_{90/95/90}</math></b>	<b>0.254</b>

### Signal-Response NDE Method Example

A Standard NDE POD study for a signal-response NDE method is illustrated with 40 flaws and 10 certified inspectors randomly sampled from a specific population of interest. Every inspector sees each of the 40 flaws once and reports the signal response from the NDE method in a fully crossed design. The scan and/or waveform is recorded, but it is not considered in this simplified example. POD of this NDE method is assumed to be related to a single flaw size characteristic (e.g., flaw depth). A decision threshold ( $y_{dec}$ ) of 600 and a lower (left) censoring signal of 200 were derived prior to the Standard NDE POD study. Table B.3 shows the inspection data from the flawed specimens. The flaw specimens are assigned a unique, noninformative two-letter sequence unrelated to flaw size to avoid prompting the inspector to guess at the specimen flawed or unflawed condition. The flaw sizes were independently characterized using computed topography (CT). This example was inspired by MIL-HDBK-1823A (2009), Appendix G, Example 1 dataset, and it is not intended to represent a particular NDE method.

To estimate the POF, the 10 inspectors are presented with 40 unflawed sites and signals (i.e., noise signals) are reported from the NDE method. The unflawed inspection data are presented in Table B.4. In practice the 40 flawed and 40 unflawed specimens/sites would be randomly intermingled and presented to the inspectors sequentially. However, the flawed and unflawed inspection signals are shown in separate tables for clarity in illustrating the POD and POF analysis.

The 1/50 POF signal level for each inspector is the estimated upper tolerance bound for 99% coverage (i.e., 1% POF) with 50% confidence from a lognormal distribution, and they are shown in Table B.5. NASA/TM-20210018515 (Koshti, et al. (2021)) discusses the POF analysis approach used, albeit for 1/95 POF. By inspection, none of the POF signal levels exceed the decision

threshold ( $y_{dec}$ ) of 600. Therefore, all of the inspectors are included in the estimation of the Standard NDE flaw size.

Figure B.4 shows the signal response versus the flaw size for the inspectors from the data contained in Table B.3 for the flawed specimens. A linear model,  $y = \beta_0 + \beta_1 x + \varepsilon$ , is fit for each inspector, where  $y$  is the signal,  $x$  is the flaw size,  $\beta_0$  is the intercept,  $\beta_1$  is the slope, and  $\varepsilon$  is the random error that is assumed to be an independently, identically distributed, normal random variable with a mean of 0 and a variance of  $\sigma^2$ . Censored regression is performed using a lower (left) censoring signal of 200, and the model parameters are estimated as  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $s$  for each inspector. The difference among the linear lines illustrates inspector-to-inspector variability. The flaw design features 10 nominal flaw sizes with 4 replicates of each size, which can be seen by vertical groupings with some horizontal variability due to the variation in the production of the actual flaw sizes.

Table B.6 provides the estimated model parameters for each of the individual inspectors, and the a90/95 flaw size for each inspector is computed as based on the 90/95  $k_1$  value of 1.710 for 38 degrees of freedom in the residual error (i.e., 40 flaws minus 2 degrees of freedom for estimating the regression parameters).

$$\widehat{a_{90/95}} = \frac{1}{\hat{\beta}_1} (y_{dec} - \hat{\beta}_0 + k_1 * s)$$

The average a90/95 is 0.109, and the standard deviation of the estimated a90/95 values among the inspectors is 0.013. For 10 inspectors, the  $k_1$  value for the 90% percent of the inspectors with 50% confidence is retrieved from the table in Appendix A as 1.324. Using Equation 3 provided in Section 5.3, the Standard NDE flaw size is estimated to be 0.126, which is the a90/95/90 Standard NDE flaw size. In this example, the Standard NDE flaw size is smaller than the largest individual inspector a90/95 of 0.141. This outcome illustrates the 90% inspector coverage of the estimated Standard NDE flaw size.

Table B.3. Signal-Response Standard NDE POD Example Dataset with 40 Flaws and 10 Inspectors

Index	Specimen ID	Flaw Size	Reported Signal from each Inspector									
			1	2	3	4	5	6	7	8	9	10
1	GR	0.021	267.8	342.9	742.1	651.6	452.7	298.4	310.4	398.0	264.7	208.3
2	AN	0.019	564.2	144.5	43.3	260.4	308.1	279.5	486.9	293.2	66.2	215.5
3	GI	0.019	442.3	206.9	83.1	395.6	0.8	266.4	272.2	411.1	208.3	362.6
4	BL	0.024	412.2	392.0	397.7	508.9	108.4	540.1	292.7	475.7	584.0	408.0
5	EJ	0.039	324.1	428.7	383.8	66.1	579.9	594.5	481.6	433.6	416.0	250.2
6	GL	0.038	473.3	418.7	280.9	640.6	526.3	346.3	648.2	749.0	428.6	471.5
7	BA	0.040	540.8	527.7	351.8	357.4	603.3	573.0	658.0	306.8	686.4	552.0
8	FM	0.038	346.8	377.5	449.2	340.2	541.5	269.4	513.3	429.0	379.3	604.7
9	DO	0.062	412.5	652.2	477.0	599.9	557.5	507.0	616.7	692.2	665.9	472.9
10	GH	0.062	479.7	453.7	690.0	480.3	709.6	623.7	690.5	534.3	472.6	685.4
11	AU	0.059	469.7	781.1	919.9	760.5	702.8	594.7	672.2	358.3	477.2	557.1
12	CG	0.058	654.7	420.2	632.7	739.5	586.5	471.3	419.8	535.2	652.6	632.9
13	EL	0.079	632.5	718.4	601.0	973.7	410.8	683.8	604.6	857.6	768.6	547.5
14	ED	0.080	554.6	633.5	348.2	910.9	598.8	680.2	552.5	494.1	724.4	705.9
15	CM	0.079	500.3	312.1	467.3	805.2	647.4	694.7	734.8	770.5	622.5	577.4
16	EW	0.082	666.1	634.9	781.0	646.3	570.8	699.3	750.8	371.8	831.4	729.0
17	DP	0.093	896.6	872.0	503.8	707.4	737.5	745.7	922.1	627.1	787.0	764.2
18	CE	0.100	675.5	652.5	761.5	836.0	907.2	797.5	1025.2	862.9	744.7	923.3
19	DT	0.101	697.5	917.6	995.1	723.5	787.7	621.7	686.7	815.2	634.8	725.7
20	BE	0.097	606.6	742.6	698.8	765.2	683.9	824.6	791.6	673.1	950.3	692.4
21	GD	0.120	874.2	1143.0	939.6	977.7	718.2	901.1	744.1	609.5	823.9	858.0
22	FQ	0.117	766.4	807.0	838.7	800.7	699.8	769.9	986.6	847.1	988.9	836.0
23	CV	0.122	944.9	1136.2	1223.7	741.1	896.8	957.2	890.3	803.3	769.5	990.0
24	BG	0.122	779.7	825.3	1117.9	750.4	849.6	977.6	985.5	843.7	827.7	795.7
25	ET	0.139	994.1	1081.0	923.8	949.4	1132.1	1008.5	1121.6	1247.5	941.1	899.0
26	GJ	0.139	924.0	1010.2	1133.4	961.1	1036.3	872.5	1075.4	958.9	1122.0	746.4
27	GK	0.140	864.0	1061.4	860.6	1048.6	1095.5	819.5	1077.5	967.0	995.1	937.9
28	CK	0.143	924.2	1028.9	1040.0	1061.3	922.0	798.5	1161.8	903.4	941.6	919.0
29	AI	0.161	969.3	1267.2	626.8	1027.7	1143.1	1112.0	1241.5	942.2	1080.1	1106.0
30	DV	0.157	1089.4	1229.8	828.7	1152.2	1117.3	994.6	1101.9	1215.1	1024.6	1005.9
31	FO	0.160	967.1	1062.5	988.1	1070.2	1277.4	1042.9	1116.3	1227.6	1052.0	951.4
32	BT	0.160	985.3	1131.4	927.1	1117.4	1055.9	894.9	1081.3	1167.2	1069.9	1040.8
33	AT	0.184	1023.6	1246.0	1206.6	1180.0	997.5	1051.3	1062.1	1157.0	1156.8	1194.5
34	EY	0.177	1211.0	1158.5	844.9	1143.6	1135.4	1159.2	1216.7	1272.7	1142.1	1027.2
35	GE	0.178	1175.1	1103.1	1004.7	1112.1	821.9	1044.7	1032.2	1254.0	1161.6	935.0
36	BF	0.182	1095.2	1089.2	1165.1	1095.8	1258.4	978.9	1374.2	1191.0	1131.5	1144.5
37	AV	0.202	1240.1	1308.1	1205.0	1378.8	1356.5	1369.5	1379.8	1100.1	1476.4	1230.2
38	CI	0.198	1205.3	1319.6	1434.2	1314.6	1534.1	1285.2	1468.3	1039.7	1203.3	1328.4
39	DL	0.204	1148.7	1180.1	993.5	1099.1	1352.5	1055.7	1557.4	1305.1	1400.7	1349.9
40	AP	0.198	1266.2	1256.8	1180.9	1031.8	1294.2	1148.1	1303.2	1275.6	1091.0	1166.7

Note: Flawed specimen data only is shown. However, 40 unflawed specimens/sites were intermingled with the flawed specimens during the inspection to estimate the POF.

Table B.4. Signals from 40 Unflawed Sites to Estimate POF

Index	Specimen ID	Inspector Reported Signals (Noise Signals) from Unflawed Specimen									
		1	2	3	4	5	6	7	8	9	10
41	AB	88.4	110.7	60.5	106.9	54.8	60.3	70.1	90.5	121.2	80.4
42	GT	72.3	143.2	72.0	137.0	81.7	179.5	86.3	164.6	103.8	85.3
43	EQ	71.3	143.1	113.7	100.4	137.7	50.4	87.1	102.2	99.9	108.8
44	FD	111.0	93.7	110.1	149.8	100.1	87.3	99.8	101.0	141.0	137.0
45	GF	123.1	233.1	150.7	181.5	72.8	73.7	161.8	121.5	79.8	95.1
46	FV	139.7	136.6	104.3	86.2	160.5	83.8	145.6	130.5	122.7	81.4
47	AY	57.0	188.4	70.4	102.1	162.0	133.9	61.9	111.8	52.1	110.2
48	CD	109.2	96.5	273.3	56.7	41.1	117.1	102.3	131.7	114.7	118.8
49	AJ	152.0	125.6	159.7	84.3	149.0	66.8	79.2	121.1	203.7	124.8
50	FS	65.0	64.1	127.2	134.1	98.7	131.7	117.7	62.1	83.3	258.6
51	CA	36.4	112.9	196.9	129.3	148.3	148.1	65.0	122.2	207.6	111.2
52	DZ	154.5	118.6	113.0	143.1	63.2	137.0	85.6	86.9	100.5	87.8
53	FZ	87.1	138.3	60.4	173.7	47.7	113.2	80.2	57.0	99.8	70.7
54	CF	90.2	85.1	140.6	96.7	176.1	86.8	124.6	146.5	87.6	118.1
55	GB	185.8	88.3	101.8	146.1	118.9	59.5	65.7	131.5	115.3	167.4
56	BK	75.0	123.1	119.7	146.0	176.5	65.6	122.0	106.9	176.3	49.0
57	BP	155.2	93.0	174.9	107.1	83.6	80.8	68.2	118.4	117.4	70.3
58	CZ	70.3	171.8	122.0	140.8	104.3	188.8	70.4	139.6	311.3	60.2
59	CX	68.1	111.7	213.3	52.8	105.2	83.5	160.1	115.4	89.5	78.3
60	GW	181.2	62.3	88.9	68.1	57.7	60.5	45.7	115.6	110.7	155.3
61	DF	196.1	126.3	122.6	79.6	97.8	47.1	100.5	49.2	218.8	35.0
62	AQ	110.8	197.5	130.3	60.8	138.1	74.1	180.1	73.2	137.6	90.1
63	GG	80.9	95.7	91.1	147.2	73.0	156.6	131.6	79.1	83.4	120.0
64	FG	107.4	76.2	138.8	115.2	45.8	143.3	185.5	206.2	142.1	108.6
65	EO	62.8	119.6	70.1	217.6	61.7	86.2	93.0	116.5	106.7	176.0
66	FH	97.2	126.7	61.0	55.7	110.7	67.2	184.5	82.1	103.8	82.3
67	BI	84.6	66.9	171.6	212.7	103.9	150.8	138.6	87.7	56.0	287.8
68	FN	176.4	83.5	164.8	161.4	55.8	150.1	57.0	99.2	95.1	74.5
69	EV	152.4	110.4	123.9	97.3	65.4	61.6	85.8	101.8	69.6	48.4
70	EZ	69.7	94.7	108.8	76.3	113.9	136.0	57.7	124.6	150.1	163.8
71	EC	78.8	69.3	145.0	200.9	105.1	126.8	69.1	112.4	156.9	126.1
72	ER	89.9	127.3	126.8	55.9	113.7	146.4	199.6	77.0	89.9	133.3
73	DM	95.1	51.4	84.5	194.5	71.2	157.6	100.1	53.6	78.4	146.6
74	EM	108.2	76.1	48.4	170.6	101.6	116.6	214.3	86.9	34.5	70.6
75	ES	69.0	185.2	81.7	56.7	81.1	297.9	114.0	110.0	110.0	61.2
76	EA	146.1	70.3	66.3	174.9	81.6	38.9	74.0	100.5	112.0	185.5
77	FA	113.5	86.4	88.0	63.8	97.7	130.1	209.6	87.1	112.1	95.5
78	FU	87.7	88.7	188.2	67.9	116.8	178.8	151.7	183.4	129.5	103.6
79	DG	164.9	55.8	147.2	115.7	33.7	135.3	104.3	62.2	115.0	174.2
80	BO	163.0	137.9	109.1	47.0	144.2	94.5	96.5	123.4	115.1	149.5

Table B.5. Estimated 1/50 POF Signal for Each Inspector

	1/50 POF
Inspector	Signal
1	251.4
2	243.5
3	281.2
4	305.8
5	242.3
6	293.0
7	265.3
8	215.7
9	279.4
10	296.6

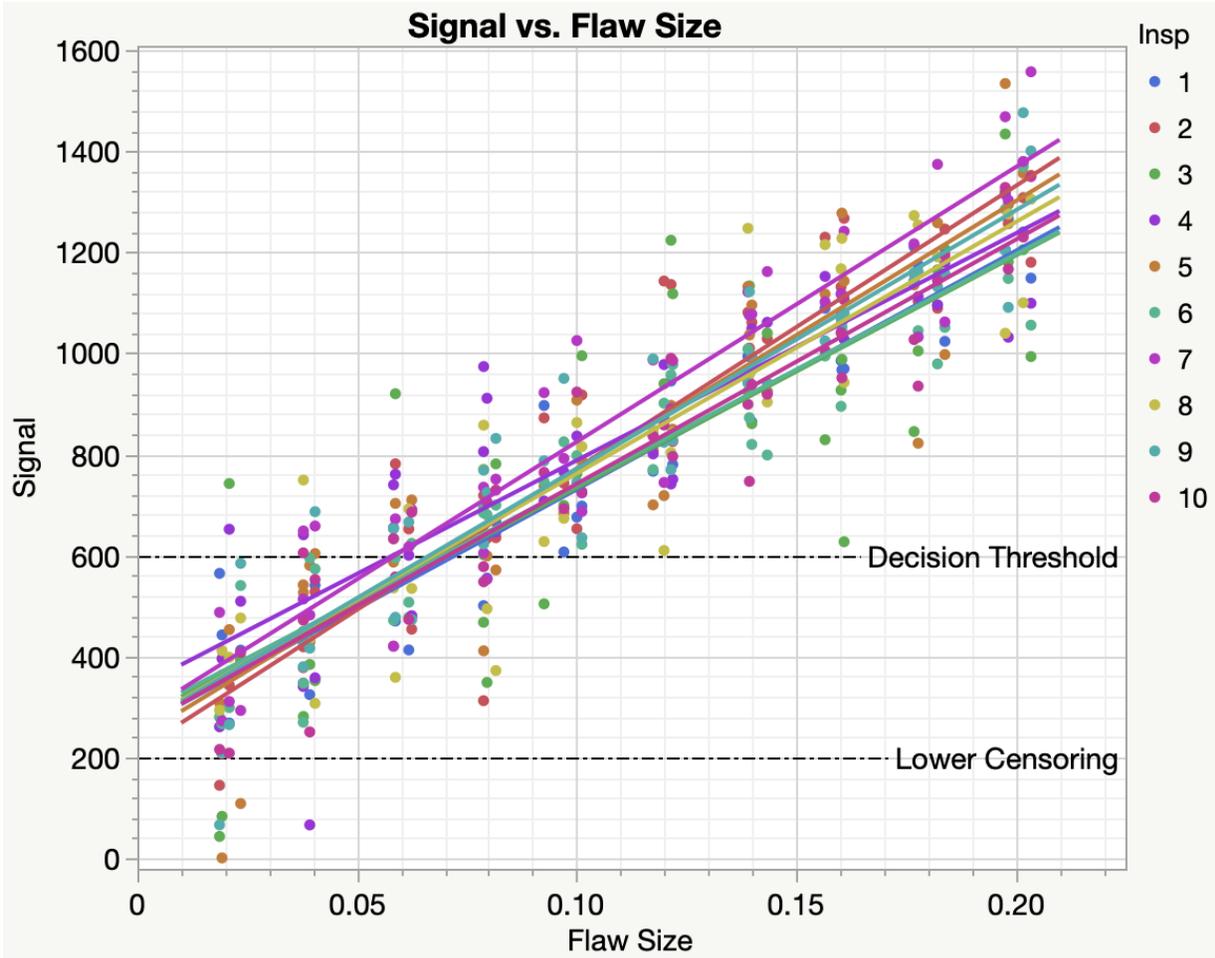


Figure B.4. Individual inspector signal versus flaw size models for a signal-response Standard NDE POD example.

Table B.6. Individual Inspector Model Estimated Parameters, a90/95 flaw size, and the Standard NDE flaw size, a90/95/90, for the Signal-Response NDE Method Example

		Decision Threshold Signal	600
		Lower Censoring Signal	200
		k1 value, 90/95, 40 flaws (38 degrees of freedom)	1.710
<b>Inspector</b>	<b><math>\beta_0\_hat</math></b>	<b><math>\beta_1\_hat</math></b>	<b>s</b>
1	257.9	4724.2	87.7
2	212.2	5595.5	121.7
3	274.1	4597.1	189.2
4	338.3	4494.7	129.6
5	238.1	5316.6	132.3
6	282.0	4570.2	96.7
7	280.2	5441.9	110.2
8	263.6	4979.9	134.9
9	259.2	5118.6	107.5
10	257.7	4835.6	93.9
		Inspector Average a90/95	0.109
		Inspector-to-inspector variability, Stdev of a90/95	0.013
		k1 value for 0.90/50% for 10 inspectors	1.324
		<b>Standard NDE Flaw Size, a90/95/90</b>	<b>0.126</b>

## Appendix C – Inspector Sampling Discussion and Multiple Facility Guidance

In this Appendix, additional motivation and analysis of the proposed sampling plan are discussed. Spencer (2020a) recommends the number of inspectors based on the number of facilities and number of inspectors within each facility. Spencer’s sampling strategy provides an approach to independently estimate facility-to-facility variation from inspector-to-inspector variation between and within different facilities. His guidance is based on a subjectively chosen number of inspectors that provide diminishing returns in estimation precision by plotting the multiplicative factor used to compute the statistical tolerance interval versus the number of observations. The ‘knee’ in the curve was identified by Spencer to be approximately 7 observations. This is similar to the analysis performed in NASA/TM-20210018515 (Koshti et al. (2021)), where a more conservative sample size of 10 was chosen.

Based on this minimum sample size and desire to isolate facility-to-facility and inspector-to-inspector variability, Spencer (2020a) recommends that if there are  $\leq 7$  facilities, then inspectors should be chosen from each facility. For facilities with  $\leq 7$  inspectors, Spencer recommends including all of the inspectors from that facility. For facilities with  $> 7$  inspectors, Spencer recommends a random sample of 7 inspectors be chosen. If there are more than 7 facilities, then a random sampling from these facilities would be selected with the probability of a given facility being included in the sample being proportional to the number of inspectors in that facility, where facilities with more inspectors are more likely to be selected. After the 7 facilities are chosen, then random sampling of inspectors within each facility is performed. Figure C.1 provides a flow chart of the sampling plan proposed by Spencer (2020a).

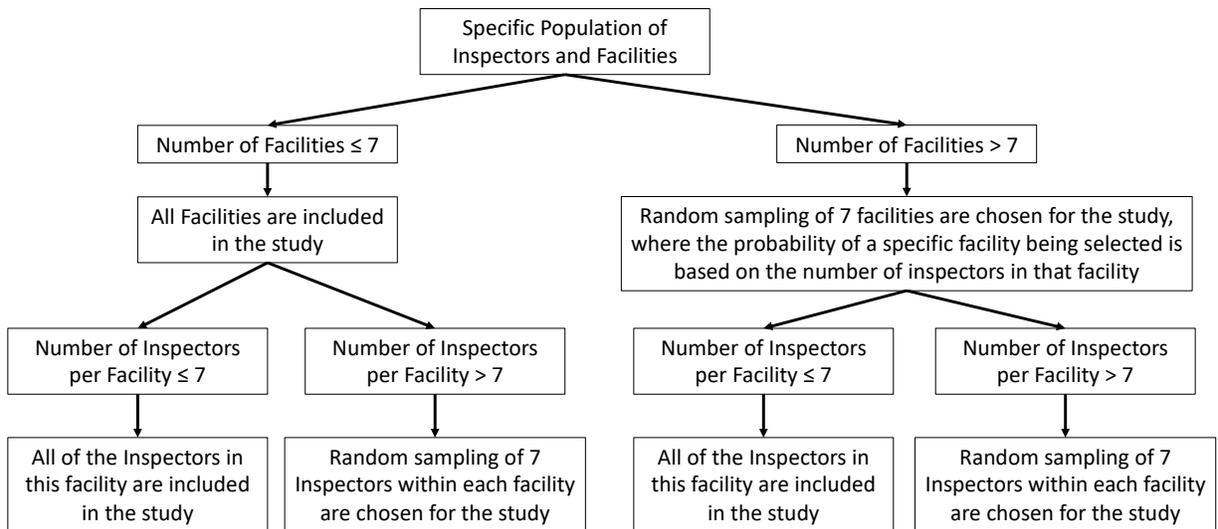


Figure C.1. Flow chart of inspector and facility sampling guidance.

Applying Spencer’s recommendation to the Bishop (1973) study, where there were 3 facilities (i.e.,  $< 7$  facilities), inspectors from each of the 3 facilities would be included in the study. While the number of inspectors within each facility was not documented in Bishop (1973), for the purpose of illustration assume that these 3 facilities each had 10 inspectors. Therefore, 7 of the 10 inspectors would be chosen at random from the 3 facilities. This results in a total of  $3 \times 7 = 21$  inspectors in the Standard POD study. Alternatively, if it is assumed that the 3 facilities each had 3 inspectors (i.e.,  $< 7$  facilities and  $< 7$  inspectors per facility), then  $3 \times 3 = 9$  inspectors would be included in

the Standard POD study, which includes every inspector. Note that a Standard NDE study with every inspector that will perform field inspections is similar to a Special NDE study, except the Standard NDE study assumes that the inspectors represent current and future inspectors with similar certifications.

While Spencer's sampling structure offers helpful insights, it is assumed that most NASA programs will have a small number of facilities and few inspectors per facility, which is different from larger operational inspection organizations (e.g., commercial airlines or the USAF) that tend to have many inspectors located at many facilities. In addition, for the purpose of simplification, it is assumed that independently estimating facility-to-facility and inspector-to-inspector variability is not an objective in a Standard NDE study.

In a limiting example, a Standard NDE study with 1 facility with 1 inspector is equivalent to a traditional MIL-HDBK-1823A POD study, even though it adheres to Spencer's recommendations. Even if there were 2 facilities and 2 inspectors at each, resulting in a total of 4 inspectors in the population of interest, the statistical small sample penalties would make a Standard NDE analysis approach overly conservative, and it would be difficult to defend the assumption that these 4 inspectors are representative of the future inspectors. In this case, Special NDE demonstration may be deemed more appropriate.

## **Appendix D – Alternative Statistical Approaches to Estimate the Standard NDE Flaw Size**

In this Appendix, alternative statistical modeling approaches are discussed for analyzing a Standard NDE POD study. The recommended two-stage approach of analyzing the average and variability among individual inspector a90/95 flaw sizes is considered the most intuitive and approachable to practicing NDE engineers. However, it is acknowledged that more sophisticated approaches may be considered in specific applications.

As a more statistically rigorous alternative, Spencer (2020b) proposed a modified two-stage approach that considers the variability among individual inspector's a90 flaw size (i.e., not a90/95). It adds a 95% statistical confidence interval to the 90% probability of detection of a 90% proportion of inspectors to estimate the Standard NDE flaw size. The primary argument for this method is that the a90 is the unknown statistical model parameter being estimated (i.e., not a90/95). While the argument is technically accurate, individual inspector capability is commonly considered to be based on a90/95 (i.e., not a90). Therefore, while this alternative approach may be considered statistically formulated, the recommended approach in this Technical Memorandum considers the variability in the a90/95 flaw size as the commonly accepted quantity that represents individual inspector capability.

As an alternative one-stage approach, a more generalized linear mixed model (GLMM) could be employed, where the term “mixed” indicates that there are fixed and random parameters in the model. This approach can appropriately represent the random sampling of the inspectors from the specific population of interest in a single POD model. While this is a statistically elegant approach, it represents a significant increase in modeling complexity even for experienced statisticians, and requires more advanced techniques not commonly found in commercially available statistical software packages. The estimation approach could be frequentist or Bayesian for this random parameter model, as illustrated in Koh and Meeker (2017). GLMMs are considered a significant step in complexity beyond the generalized linear models used in MIL-HDBK-1823A (2009).

MIL-HDBK-1823A (2009), Section 6.4 mentions multiple inspector studies in an effort to address a common misconception that repeated inspections of the same flaw violates statistical assumptions of independence and does not provide useful information. The apparent motivation for an inspector to perform repeated inspections of the same flaw is to artificially increase the sample size of the POD study by producing more inspection responses with a smaller number of flawed specimens to ultimately reduce the cost of producing the specimens. MIL-HDBK-1823A (2009) refers to multiple inspector studies as “repeated measures,” and offers an analysis approach in Appendix G.4.4, which is an approximate method that requires additional restrictions to be met in the dataset, and it is not recommended. While MIL-HDBK-1823A's repeated measures discussion appears to be applicable to NASA Standard NDE, the analysis approach recommended in this guidebook analyzes each inspector individually to estimate their respective a90/95 flaw size. In Section 3.3, it is suggested that it might be useful to present a flawed specimen to an inspector multiple times to assess inspector consistency. If this aspect were implemented, then these repeated inspections of the same flaw would be considered correlated in the modeling, or they may be handled as a separate analysis removing them from the POD modeling data set for diagnostic purposes. Overall, MIL-HDBK-1823A (2009), Section 6.4 provides a helpful warning to protect against inappropriate analysis. However, this Section is not directly related to NASA Standard NDE.

Standard NDE design and analysis could be cast into a gauge repeatability and reproducibility (Gauge R&R) framework, which is an established industrial statistics approach to evaluate an instrument's accuracy and operator influence in measurement system analysis (see Burdick, et al. (2005)). In industrial applications, reproducibility is related to the variation in the measurement procedure (i.e., the variability among operators using the same instrument/gauge). Repeatability is related to the variation in the measuring device. Adapting these definitions to Standard NDE, reproducibility would be related to the variability among the inspectors using the same NDE method, and repeatability would be related to the variability of an inspector in detecting flaws, which includes flaw-to-flaw variability (i.e., reproducibility is related to between-inspector variability, and repeatability is related to within-inspector variability). A Gauge R&R study is a designed experiment that requires multiple operators to perform inspections on the same collection of parts, suggesting its conceptual applicability to Standard NDE. The objective of the study is to isolate and estimate individual sources of variability through an analysis of variance of fixed and random effects related to the measurement process. Differing from the Standard NDE context, Gauge R&R studies typically involve the recording and analysis of direct measurements on individual parts (e.g., length of a part with a dial caliper), rather than a derived characteristic (e.g., an individual inspector's a90/95 flaw size) by modeling the calls from the entire collection of flaws. In addition, Gauge R&R studies are usually designed to ensure that a measurement system meets a pre-specified accuracy rather than the objective to estimate the detection capability of an NDE method. Notwithstanding these differences in the analysis process and objective, the Gauge R&R discipline provides a useful perspective for Standard NDE studies.

These alternative approaches offer viable alternatives to the recommended two-stage approach of analyzing the average and variability among individual inspector a90/95 flaw sizes.

## **Appendix E – Checklist of Guidance for the Design and Analysis of a Standard NDE Study**

### **Summary of Standard NDE Study Expectations**

Based on the proposed requirement for NASA-STD-5009C in Section 2.0.

A Standard nondestructive evaluation (NDE) probability of detection (POD) study consists of a MIL-HDBK-1823A (2009) compliant POD study that is conducted by a minimum of 10 inspectors that are a representative sample from a specific population of inspectors.

Individual inspector analyses are performed in accordance with MIL-HDBK-1823A methods, and the estimated a90/95 flaw sizes for the individual inspectors are reported. Individual inspector probability of false calls (POF) are reported and do not exceed 1% POF with 50% confidence.

The Standard NDE flaw size is estimated as a function of the average and standard deviation of individual inspector a90/95 flaw sizes and represents the flaw size that 90% of inspectors are expected to demonstrate at least 90/95 detection capability.

### **Checklist of Standard NDE Study Guidance**

Based on guidance provided in referenced sections.

#### **1.0 Introduction (page 1)**

1. Statistician is consulted in the planning, analysis, and reporting (MIL-HDBK-1823A (2009), Section 4.5.1.b.).

#### **3.0 Standard NDE Study Design (page 3)**

2. Study design is guided by the intended application(s) scenarios of the resultant Standard NDE flaw sizes and considers similarity and transferability to envisioned flight components.

#### **3.1 NDE Method Specifications (page 5)**

3. Study is designed for a single NDE method's expected detection capability, physics, specimen characteristics, and flaw type. A flawed specimen set is specifically designed or chosen for the NDE method under study.
4. Signal decision threshold derivation is documented and is consistently utilized by all inspectors and facilities for a signal-response NDE method.
5. Data recording protocol supports traceability and reproducibility of the study. Raw inspection signals are recorded and archived in addition to the inspector's call. Inspection images for an image- or scan-based hit/miss method (e.g., radiographic) are recorded, preferably in a digital format. Indicated flaw location on the specimen are recorded.
6. Inspector training is documented to ensure a consistent NDE technique representative of operational field inspections.

#### **3.2 Specimen Characteristics (page 6)**

7. Specimen geometry, material, and flaws are representative or conservative relative to field inspections.
8. Naturally occurring or simulated induced flaws (e.g., fatigue cracks) are used to provide representative flaw-to-flaw variability.

9. Crack morphology of induced flaws (e.g., aspect ratio and crack opening) are assessed as being representative of or conservative to naturally occurring flaws by a materials engineer.
10. Method to induce flaws mimics the fabrication and/or operational usage that is expected to produce naturally occurring flaws. Flaw production technique does not influence the NDE method's detection capability. Detailed technical documentation of specimen and flaw production are recorded.
11. Flaw locations are not easily deduced by the inspector for each specimen (e.g., the flaw should not always be near the center of the specimen).

### **3.3 Statistical Flaw Size Design (page 7)**

12. Prior knowledge from previous POD studies is leveraged to the greatest extent possible in the flaw size design.
13. Flaw size distribution spans from rarely detectable (POD near 0) to consistently detectable (POD approaching 1).
14. For signal-response NDE methods, maximum flaw size avoids saturation of the signal that occurs when a further increase in flaw size does not result in an increase in signal. Maximum flaw size does not greatly exceed the flaw size associated with the signal decision threshold. Minimum flaw size is below the decision threshold for a signal-response method.
15. For hit/miss NDE method, maximum flaw size is  $\approx a_{97}$  (POD of 97%) and the minimum flaw size is  $\approx a_3$  (POD of 3%) (Annis et al. (2012) Section 6.2).
16. Flaw sizes are uniformly spaced between the maximum and minimum flaw sizes (MIL-HDBK-1823A (2009), Section 4.5.2.2.a). If a transformation of the flaw size is used (e.g.,  $\log(\text{size})$ ), the distribution of flaw sizes are uniform based on the transformed flaw size (Annis et al. (2012) Section 6.2).
17. Approximately 50% of the study's flaws are in the transition region, a metric known as 'overlap.' Percentage of misses is between 30% to 50% of the total number of inspection calls, known as 'evenness' (Henry et al. (2022)).
18. Replicated flaws of the same nominal size are included throughout the range of flaw sizes.
19. Minimum of 40 flaw specimens are used for a signal-response method (MIL-HDBK-1823A (2009), Section 4.5.2.2.b).
20. Minimum of 60 flaw specimens are used for a hit/miss method (MIL-HDBK-1823A (2009), Section 4.5.2.2.b).
21. Minimum of 40 unflawed specimens/sites are used for a signal-response method, and minimum of 60 unflawed specimens/sites are used for a hit/miss method.

### **3.4 Statistical Inspector Sampling Plan (page 12)**

22. Random sampling of inspectors from the specific population of interest is representative and unbiased.
23. Minimum of 10 inspectors are included in the study.
24. Inspector selection procedures and rationale for the inspector sampling plan are documented.

### **3.5 Independent Flaw Size Characterization (page 14)**

25. Method for independent characterization of the flaw sizes is specified in the planning stage of the POD study.
26. Every flaw is independently characterized.

### **4.0 Execution (page 15)**

27. Execution protocol is documented and independently monitored.
28. A test monitor is designated to assure the execution protocol is followed (MIL-HDBK-1823A (2009), Section 4.5.3). Test monitor is present at each facility during the inspection process.
29. Briefings are developed to provide consistent instructions to the inspectors and/or facilities. A “dry run” is performed with a limited number of inspectors to ensure the briefings are adequate and/or identify omissions in the instructions.
30. Initial specimen inspection is conducted with photographic documentation to establish a baseline condition for future reference and revalidation.
31. Cleaning materials and protocol do not damage the specimen or influence detection capability. Intermittent physical inspections are performed to detect changes that might affect detection capability. Specimen cleaning is performed after every inspection. Specimen/flaw degradation is noted in the periodic inspections.
32. 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.
33. Training set of specimens is designated for technique development and practice (MIL-HDBK-1823A (2009), Section 4.5.2.c.1).
34. Inspections of the primary specimen set are performed in a blind manner (i.e., the inspector has no knowledge of whether it is a flawed or unflawed specimen/sites, nor does the inspector have an indication of the flaw location on the specimen).
35. Every inspector performs an inspection on every specimen in a fully crossed design.
36. Custom-designed shipping containers have a designated location for each specimen. Specimens are protected from mechanical damage or contamination during shipment. Pre- and post-shipment inspection and documentation are compared to the baseline physical inspection to ensure that no damage or wear has occurred.
37. Noninformative specimen designations are assigned randomly. Specimen markings and designation are not indicative of the specimen characteristics (e.g., whether a flaw is present, its size, or location) and are not correlated with any flaw characteristic (e.g., flaw size increasing with the specimen number).
38. Association of the unique specimen identifier with the specimen characteristics (e.g., flaw size) is only available to the test monitor, proctor, NDE engineer, and/or statistician overseeing the study.
39. Inspectors are presented the flawed and unflawed specimens/sites in a randomized order to preclude the ability to detect a pattern that might lead to inspector familiarity or guesses regarding flaw presence. Each inspector is presented the specimens in a unique randomized order.

40. Inspection grid locations marked on the specimen are representative of the NDE method's detection capability and limitations in field inspections.
41. Inspector fatigue due to sequential inspections is considered in the execution protocol. Inspection duration is consistent and representative of the expected operational field inspections.
42. Destructive flaw characterization is performed after a preliminary assessment of the inspection data quality to investigate anomalous inspection results.

## **5.0 Analysis (page 17)**

### **5.1 Estimating Individual Inspector a90/95 (page 19)**

43. Consistent POD modeling approach (i.e., flaw and/or signal transformations and link function) is used for all inspectors, rather than unique model specifications for each inspector.

### **5.2 Estimating Individual Inspector Probability of a False Positive (page 19)**

44. Average and standard deviation of the unflawed specimen/site signals are used to estimate the POF based on the signal decision threshold for a signal-response method.
45. Count of false positive indications are used to estimate POF for a hit/miss method.
46. POF is reported for each inspector, and every inspector demonstrates a maximum of 1% POF with 50% confidence, a 1/50 POF.

### **5.3 Estimating Standard NDE Flaw Size (page 20)**

47. 90% inspector coverage at 50% confidence based on individual inspector a90/95 flaw sizes defines the Standard NDE flaw size in the absence of specific requirements.
48. If a logarithm transformation is applied to the flaw size in the individual inspector POD modeling, then the variability of a90/95 flaw sizes is estimated with the untransformed flaw size (i.e., in the original units).
49. The distribution of individual inspector a90/95 flaw sizes was examined to identify anomalously small or large detectable flaw sizes and/or distinct clusters of flaw sizes that may warrant further investigation.
50. Analysis and modeling details are recorded to allow the results to be independently, numerically reproduced.

## **6.0 Documentation (page 21)**

51. Comprehensive inspection report and a condensed summary specification are generated for the Standard NDE study that are sufficient to allow independent reproduction of the results.