NASA/TM–20210018515/Corrected Copy
NESC-TI-20-01545

Guidebook for Limited Sample Probability of Detection (LS-POD) Demonstration for Signal-Response Nondestructive Evaluation (NDE) Methods

Ajay Koshti
Johnson Space Center, Houston, Texas

Peter A. Parker
Langley Research Center, Hampton, Virginia

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

July 2021

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

- 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–20210018515/Corrected Copy
NESC-TI-20-01545

Guidebook for Limited Sample Probability of Detection (LS-POD) Demonstration for Signal-Response Nondestructive Evaluation (NDE) Methods

Ajay Koshti
Johnson Space Center, Houston, Texas

Peter A. Parker
Langley Research Center, Hampton, Virginia

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

July 2021

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 for performing limited sample probability of detection (LS-POD) demonstration testing with smaller numbers of flawed specimens than the 29 required for the Point Estimate POD method referenced in NASA Standard 5009B. In particular, the demonstration testing will determine whether a nondestructive evaluation (NDE) method and specific inspector (i.e., method/inspector) using a signal decision threshold for a single target flaw size will provide a POD that exceeds the required 90-percent POD with 95 percent confidence. Additionally, the demonstration testing evaluates whether the probability of false calls (POF) is less than the required value established for the given inspection (examples are provided for a recommended POF value of 1 percent with 95 percent confidence in this document). It also provides process control guidance for assuring that the field applications of the NDE procedure maintains the capability demonstrated in the POD testing.

Scope – This document is specifically applicable to NASA programs and projects where NASA Special NDE methods are used to inspect fracture critical human spaceflight metallic hardware when the minimum number of flawed specimens required for the Point Estimate POD method (i.e., 29) referenced in NASA Standard 5009B (2019) are not available. However, NASA programs and projects that are not human rated may choose to use these methods and the methods may be applicable to other material/flaw types. The guide is restricted to NDE methods that provide signal response data (i.e., not applicable to NDE methods that provide only a hit-miss response) for which the signal is a function of the flaw size. Example applicable NDE techniques include ultrasonic and eddy current inspection methods where a signal value is used as an indication of a flaw, whether the signal value comes from a single transducer or a transducer array. Similar to the Point Estimate POD method, the LS-POD method only assesses the POD/POF for a single flaw size that is designated the target flaw size.

Table of Contents

List of Figures

List of Tables

Nomenclature

EDM	Electrical Discharge Machining
LS-POD	Limited Sample Probability of Detection
NDE	Nondestructive Evaluation
NDI	Nondestructive Inspection
NDT	Nondestructive Testing
NIST	National Institute of Standards and Technology
PFID	Plasma Focused Ion Beam
POD	Probability of Detection
POF	Probability of False Calls

Definitions (adapted from NASA Standard 5009B with changes and additions specific to LS-POD method as required)

Capability Demonstration Specimens: A set of specimens made from material similar to the material of the hardware to be inspected with known flaws used to estimate the capability of indication detection, (i.e., Probability of Detection (POD) or other methods of capability assessment) of an nondestructive evaluation (NDE) method.

Cracks or Crack-Like Flaws: A discontinuity assumed to behave like a crack for assessment of material or structural integrity.

Defect: One or more flaws whose aggregate size, shape, orientation, location, or properties do not meet specified acceptance criteria and are rejectable.

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.

Hit-Miss NDE Data: Data resulting from a 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 sometimes represented numerically as a 1 (i.e., indication present) or 0 (no indication). No signal measurements from any NDE sensor output are recorded.

Initial Crack (Flaw) Size: The crack size that is assumed to exist in the part for damage tolerance analysis.

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 a 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 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 by intentional loading (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.) proportional to flaw size is measured. The determination for whether an indication is present is typically made based on a threshold value of the signal-response NDE data.

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 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.

Target Flaw Size: Flaw size that is established for which the NDE method and inspector must provide reliable (i.e., 90/95 POD) detection. This flaw size is then used in POD demonstration testing.

1.0 Introduction

NASA Standard 5009B specifies that nondestructive evaluation (NDE) methods/inspectors used to inspect fracture critical metallic components in human-rated aerospace flight systems provide 90-percent probability of detection (POD) with 95 percent confidence (i.e., 90/95 POD). This can be accomplished in two ways. The first is to use a designated Standard NDE method and the established Standard NDE flaw sizes contained within NASA Standard 5009B along with the prescribed industry standard technique processes. For all other situations, the NDE method/inspector is designated as Special NDE and currently the 90/95 POD must be demonstrated by POD testing using one of two approved methods. The first acceptable POD method is through the use of POD testing/analysis described in MIL-HDBK 1823A (2009) to provide an estimate of the minimum flaw size that meets the 90/95 requirement. Such POD testing requires a minimum of 40 to 60 NDE simulated flaws of varying sizes above and below the 90/95 POD flaw size. The second method of demonstrating 90/95 POD for Special NDE is known as the Point Estimate method as described by Rummel (1982). This method does not provide an estimate for the minimum 90/95 POD flaw size but evaluates whether the method/inspector provides 90/95 POD or greater for a specific target flaw size. For this method, a minimum of 29 specimens containing NDE simulated flaws of the target size are required. For both methods, an additional number of unflawed specimens (or unflawed inspection opportunities on the same specimens) are required to estimate the probability of false calls (POF). The intent of this guidebook is to provide a third method to demonstrate POD for NASA Special NDE applications. However, until acceptance and incorporation of this new method into NASA Standard 5009B, the use of this method will require approval by the responsible NASA Technical Authority (e.g., Fracture Control Board) on a case-by-case basis.

Although NASA Standard 5009B requires that the POF be established for every POD analysis method, it does not provide any specific requirement for the maximum POF allowed value. There are multiple considerations for establishing a maximum POF. If the POF is too high, then it can bring into question the integrity of the POD estimate (e.g., in the extreme, calling every inspection opportunity a flaw will result in 100-percent POD, but incorrectly identifies all non-flaws as flaws). Additionally, there are potential economic costs and schedule delays to falsely identifying acceptable hardware as flawed, thus resulting in complex structural analyses, additional inspections, unnecessary repairs, and/or component scrappage. Thus, minimizing POF is highly desired. However, the POD and POF are inter-related such that decreasing the POF may result in increasing the detectable flaw size that maintains 90/95 POD. Therefore, an acceptable technical and programmatic balance between POD and POF must be achieved. A key aspect to analyzing false calls is having an adequate number of inspection opportunities in which flaws are known not to exist in the POD testing specimens. An acceptable maximum allowed POF while meeting the required 90/95 POD should be established as part of the POD test plan in coordination with the program requiring the inspection and the responsible NASA Technical Authority (e.g., Fracture Control Board). For the Limited Sample Probability of Detection (LS-POD) methodology described herein, a value of 1-percent POF with 95-percent confidence (i.e., 1/95 POF) was used throughout the method description and examples provided. A table for the parameter needed to easily estimate the 1/95 POF as a function of sample size is also provided in Appendix D. The 1/95 POF criterion is similar to the MIL-HDBK-1823A guideline document, which also does not identify a specific value, but provides examples where 1-percent and 0.1-percent POF are calculated without an associated confidence value. If another POF target value is selected for use by the program/technical authority, then statistical expertise should be engaged to perform the necessary calculations as details for other target POFs are not provided in this guidebook.

For many NASA applications, it is difficult to produce the required set of 29 NDE simulated flaws necessary for NASA Special NDE point estimate POD testing. Due to this challenge, the LS-POD methodology has been developed to ensure that the required 90/95 POD is provided with the suggested 1/95 POF. The methodology is described in detail by Koshti (2021) and the statistical details are provided in Appendix A. It is conceptually based on a statistical hypothesis testing framework for decision making, where the statistical power is the POD, see Walpole, Myers, and Myers (1998). Although statistical hypothesis testing has not been recognized and applied in the NDE practice, Olin and Meeker (1996), and Swets (1983a), and Swets (1983b) suggest it for NDE applications. The methodology contained in this guidebook builds on the concepts of classical statistical power analysis to accommodate the 95-percent confidence requirement for the estimated 90-percent POD flaw size required by NASA Standard 5009B. As described in Appendix A, the approach is analogous to MIL-HDBK-1823A's estimation of the 90/95 flaw size with difference being that MIL-HDBK-1823A operates on a distribution of multiple flaw sizes, while the LS-POD approach operates on the distribution of signals from multiple specimens of a single target flaw size. However, validation experiments, described in Appendix B, with simulated and measured NDE data show that the methodology provides results conservative to MIL-HDBK-1823A for these data sets with smaller numbers of flawed signals. However, this validation is limited to signal data that meets the conditions required for this analysis, which is described in Appendix A and Section 2.6. Once the required 90/95 POD and recommended 1/95 POF have been established through demonstration testing for the target flaw size, process monitoring techniques can be implemented to ensure that the demonstrated POD/POF are valid in the hardware inspections. Detailed instructions for applying the limited flaw specimen set POD/POF methodology are provided in the guidebook. Also included are specific guidelines for using the methodologies, and critical assumptions that, if not checked, may invalidate the results from these methods.

2.0 Limited Flaw Specimen POD Instructions

Below are specific instructions and formulas for performing the calculations to demonstrate at least 90/95 POD with less than 1/95 POF.

2.1 Input Data Requirements

Two sets of NDE signal data are required for this analysis method, acquired from a set of specimens with NDE simulated flaws of the target size (i.e., capability demonstration specimens), and from specimens with no flaws or regions of the capability demonstration specimens without flaws (i.e., unflawed signal data). The signals metric (e.g., peak amplitude, peak-to-peak amplitude, etc.) from each set of specimens should be the same and measured in an identical manner. For the demonstration testing, ideally the flawed and unflawed specimens or inspection opportunities should ideally be randomized such that the inspection is performed blind with respect to knowledge of the existence or lack of a flaw for a given signal measurement. Blind testing is a requirement for other types of POD testing and should be employed when practical for LS-POD testing. This is to ensure that the inspector does not "optimize" the signal response by, for example, manipulating the sensor/couplant to get the largest signal response, when acquiring data at a known flaw location. However, blind testing is not specifically required within the analysis methodology to estimate the 90/95 POD and 1/95 POF signal limits as they are estimated independent of any specified decision threshold or inspector determination of an inspection indication. Once the 90/95 POD and 1/95 POF signal levels have been estimated, if a prior decision signal threshold for defect indication was established, then it can be evaluated to determine if it meets the POD/POF requirements (i.e., lies between the estimated 90/95 POD and 1/95 POF signal levels). Alternatively, if no prior signal threshold exists, then a value can be established from the results of the demonstration test if the condition is satisfied that the estimated 90/95 POD is greater than the estimated 1/95 POF signal levels.

2.1.1 Unflawed Signal Data

The unflawed signal data set is obtained from multiple NDE inspection measurements from unflawed specimens. This data set is used to establish the intrinsic variability of the measured signal, which in this document is referred to as "noise." However, it is noted that this needs to include all suspected sources of significant variability in the measurement that might occur in the NDE inspection and not just measurement system electronic noise. Examples include variability: in attaching/coupling the NDE sensor or measurement device; from changing environmental conditions (e.g., ambient temperature); in the specimen itself (e.g., surface roughness or variable material composition or form); and associated with the measurement system (e.g., electronic noise). A minimum of 40 such measurements exercising all sources of possible signal variability is suggested. No formal requirements are imposed as to how many samples are required for each of the identified sources of unflawed signal variability. Additionally, it is not specified how many of these measurements can be made on the same specimen (ideally in different locations) versus on different specimens. The NDE Engineer designing the test should specify the number of different measurements associated with different variables that affect the measured response in the absence of flaws. However, if known, increased numbers should be allotted to sources that to result in increased variability (e.g., if the variability from instrument electronic noise is much smaller than that due to material surface roughness variations, then the testing should include higher numbers of specimen with surface roughness variations). The unflawed signal values are designated y_{N_i} where i ranges from one to the number of unflawed signal measurements (n).

2.1.2 Flaw Signal Data

The second set of data are NDE signal measurements from a set of specimens with documented NDE simulated flaws. For the LS-POD method, the signal measurements are obtained from NDE simulated flaws that are all of the target size. These signal measurements are denoted as y_{F_i} where i ranges from one to the number of flawed signal measurements (m). Ten flawed specimens is the recommended minimum sample size based on an evaluation in Appendix B of diminishing sample size benefits in estimating the distributional parameters and NDE engineering experience. However, it is noted that the statistical uncertainty in the estimated POD signal levels is a function of the number of flawed specimen and as such the likelihood of demonstrating the required 90/95 POD may increase with increasing numbers of flawed specimen in the test. While historical estimates of the flaw-signal and noise responses are combined with economic considerations in choosing the sample size of the initial demonstration experiment, it may be necessary to sequentially specify additional measurements after the initial demonstration is performed, similar to the Point Estimate Method to achieve the 90/95 POD, or it may indicate that the target flaw size needs to be increased.

2.1.3 Decision Threshold Signal Level

For inspections, a decision threshold is required to assess whether an indication is valid. For some signal-response-based NDE methods, the decision threshold originates from other specifications or requirements relative to calibration methodology and is determined prior to the POD demonstration testing. In such cases, the demonstration testing is an evaluation of whether that threshold provides the required 90/95 POD and 1/95 POF. However, for the LS-POD method, a decision threshold is not required for the analysis and can be an output of the demonstration test results. In that case, once the 90/95 POD and 1/95 POF signal levels are estimated, and assuming that the 90/95 POD signal level is greater than that of the 1/95 POF, a decision threshold can be established for the hardware inspection within the range between and including the POD/POF bounds. Different strategies for setting the decision threshold within this range are described in Section 2.5. However, once the decision threshold is established in the initial demonstration testing (typically by the NDE Engineer or level 3 personnel) and is defined in the resulting procedure for the given inspection, the same threshold should be used in additional POD demonstration testing for any additional inspectors being qualified for Special NDE certification for that inspection.

2.2 Estimated Parameters

From the input data sets described in Section 2.1, the following input parameters are estimated to enable estimation of the POD/POF.

The mean signal in the absence of flaws ($\bar{y}_N$) is calculated by:

$$\bar{y}_N = \frac{\sum_{i=1}^{n} y_{N_i}}{n}$$

The standard deviation of noise, or signals in the absence of flaws (s_N) is calculated by:

$$s_N = \sqrt{\sum_{i=1}^{n} \frac{(y_{N_i} - \bar{y}_N)^2}{n-1}}$$

The mean signal in the presence of target flaws ($\bar{y}_F$) is calculated by:

$$\bar{y}_F = \frac{\sum_{i=1}^{m} y_{F_i}}{m}$$

The standard deviation of signals in the presence of target flaws (s_F) is calculated by:

$$s_F = \sqrt{\sum_{i=1}^{m} \frac{\left(y_{F_i} - \bar{y}_F\right)^2}{m-1}}$$

2.3 Estimation of Decision Threshold Providing 90/95 POD for a Target Flaw Size

From the Section 2.2 input data and parameters, an estimated signal value that provides 90 percent POD at 95-percent confidence is obtained from:

$$y_{90/95\ \mathrm{POD}} = \bar{y}_F - k_{1_F} * s_F$$

where k_{1_F} is the factor that gives a 95 percent lower confidence bound on the 0.1 quantile of the distribution for a specific value of m that is provided in the table in Appendix C. This is the signal level that, if selected as a decision threshold would result in a 90/95 POD for the target flaw size.

2.4 Estimation of Decision Threshold Providing 1/95 POF for a Target Flaw Size

From the Section 2.2 input data and parameters, an estimated signal value that provides 1-percent POF at 95-percent confidence is obtained from:

$$y_{1/95\ \mathrm{POF}} = \bar{y}_N + k_{1_N} * s_N$$

where k_{1_N} is the factor that gives a 95 percent upper confidence bound on the 0.99 quantile of a distribution for a specific value of n that is provided in Appendix D. This is the signal level that, if selected as a decision threshold would result in a 1/95 POF for our target flaw size.

These input parameters are shown graphically in the Figure 2.4-1.

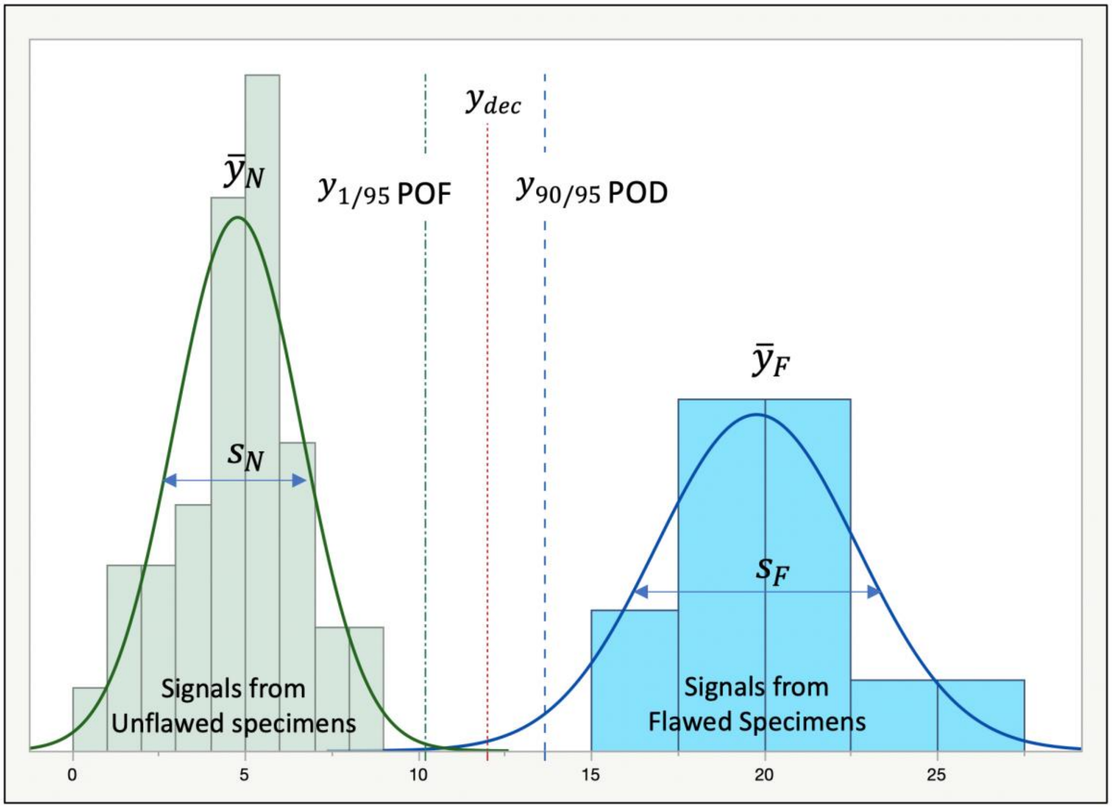

Figure 2.4-1. Graphical illustration of noise and signal distributions and resulting LS-POD analysis input parameters.

2.5 Interpretation of Results

2.5.1 Signal Decision Threshold Provided

For the case when a signal decision threshold, y_{dec}, is provided, if the results indicate that the estimated 90/95 POD signal level is greater than the estimated 1/95 POF signal level, and that the decision threshold is at or between these two levels as shown in Figure 2.4-1, such that

$y_{1/95\ \mathrm{POF}} < y_{90/95\ \mathrm{POD}}$ and $y_{1/95\ \mathrm{POF}} \leq y_{dec} \leq y_{90/95\ \mathrm{POD}}$, then the specific inspector using the specific NDE method and associated procedure demonstrated the 90/95 POD and the 1/95 POF. Thus, the NDE method/inspector is acceptable for inspection of the hardware. Per the requirements of NASA Standard 5009B for Special NDE, any additional inspectors would have to similarly demonstrate the required 90/95 POD capability (i.e., repeat inspection steps and data interpretation outlined in Sections 2.1 through 2.5). It should be recognized that there is variability in the estimated 90/95 POD and 1/95 POF signal levels that could affect the relative comparisons described, which increases with small sample sizes and is exacerbated when the noise and flaw distributions are not well separated, as described in Section 2.6 and illustrated in Figure B.3. Additional guidance is provided in Section 3 to ensure that the inspection process/inspector continues to meet the required POD/POF during the flight hardware inspections.

2.5.2 Signal Decision Threshold To Be Determined

For the case when a signal decision threshold is to be determined as a result of the POD demonstration testing, the estimated 90/95 POD must be greater than the estimated 1/95 POF signal level as shown in Figure 2.4-1. If this results from the analysis, then the decision threshold could be selected as any value between and including those bounds. However, there are three scenarios that need to be considered. To account for any variation between the POD testing and the hardware inspections that can influence the POD and POF signal levels, a decision threshold as the midpoint between the estimated 90/95 POD and 1/95 POF levels can be selected. This would provide some tolerance to not meeting POD or POF as a result of inspection differences. However, if there is higher risk associated with missing a flaw, or alternatively falsely identifying an indication, then the decision threshold can be set at either the estimated 1/95 POF or 90/95 POD levels respectively.

2.5.3 Test Failure

If the estimated 1/95 POF signal level is greater than the estimated 90/95 POD signal level, then the test failed. There is no acceptable decision threshold that will provide the 90/95 POD and 1/95 POF. In this case, the preferred option is to increase the target flaw size and perform new POD demonstration testing with the larger flaw size. This is contingent on the hardware design/use environment being able to structurally support the larger NDE flaw size for both static loads and fatigue life. Another option would be to consider a higher POF, if acceptable to the Program/Fracture Control Board. As noted previously, additional statistical expertise should be engaged to estimate the POF for values other than the 1/95 provided in this guidebook. A higher POF will increase the likelihood of increased hardware evaluation to determine acceptability (i.e., use as is or return to print) or scrap. Additionally, depending on the POF value accepted, this approach may also result in limited separation between POD and POF to address any variabilities between the demonstration testing and hardware inspections.

If the estimated 90/95 POD and 1/95 POF signal levels are similar, then additional review is warranted. Again, either an increase in target flaw size and a repeated POD demonstration test, or acceptance of a higher POF could be considered. Alternatively, if the 90/95 POD and 1/95 POF signal levels are similar and the sample size is small, an increase in the number of flawed specimens in the testing may be used to improve the results. It is possible that the 90/95 POD and 1/95 POF signal levels relative values and their separation might be improved. It is important to note though, that this option involves additional new specimens, and not repeat measurements on the exiting flawed specimens as this is not permissible. The estimated 90/95 POD signal level is expected to increase with additional specimens.

Another option to be considered when the POD demonstration test is a failure or the estimated 90/95 POD and 1/95 POF values are similar is to implement changes that result in improvements to the unflawed signal data. Examples of such changes include improvements to the NDE technique procedure or the utilization of sensors or other inspection hardware with higher sensitivity or lower noise, or improvements to the flawed specimens/hardware to be inspected such as improved surface finish or part design.

2.6 Assumptions, Checks, and Limitations on Use of LS-POD Methodology

2.6.1 Distribution Assumption

The LS-POD method assumes that the distributions of signal responses, or their transformation such as with a natural logarithm, for the flawed and unflawed specimens are Normally distributed. For the larger number of signal measurements in the noise or unflawed signal data set, methods could be employed to check the appropriateness of this assumption. However, it is more difficult/less reliable to check when the number of flawed samples is small. Thus, knowledge of the physics of the specific NDE method, and supporting data for this method for flaws with larger data sets should be used to support the assumption that the method would be expected to provide a Normal distribution in signal response for a given flaw size. If there is any reason or data that invalidates the assumption of Normal flaw or noise signal distributions, then the LS-POD method is not valid.

2.6.2 Increasing Signal Response Check

The signal response, by default, must be an increasing function with respect to flaw size. This is a similar condition for the MIL-HDBK-1823A guidelines and the Point Estimate POD method. Additional testing with larger flaw specimens may be required to be included to adequately demonstrate this assumption is met. Alternatively, prior data from the specific NDE methodology inspecting similar types of flaws/hardware could be used to confirm this assumption.

2.6.3 Signal Saturation Check

Care should also be used to ensure that the noise and flawed signal responses are on the same amplitude scale. If additional gain or attenuation is applied to signals between the two types of measurements, the raw signal responses (e.g. percent full scale) should be adjusted to compensate for this gain or attenuation in the analysis. Also, signals should be checked to ensure they are not saturated or “clipped” at some maximum value as such data censoring would also likely lead to violating the Normal signal distribution requirement. If that is observed to occur, then adjustments to the measurement process (e.g. reducing the signal gain or adding attenuation to bring the signal back within the instrument measurement dynamic range) is required. Similarly, the signals should be checked to make sure that they are properly recorded on the low range, especially in the acquisition of the noise signal data. Often, in NDE measurements, if the signals are below a certain value, they are recorded as a zero value. Doing so could result in violating the uncensored Normal signal distribution requirement. Signal levels should be measured and recorded in all cases for both flawed and noise signal measurements, regardless of how small a value is measured.

2.6.4 Test Flaw Limitation

In POD testing, it is critical that the NDE simulated flaws are representative of the naturally occurring flaws expected in the hardware (i.e., NDE simulated flaws must produce signals that are consistent with, or conservative to (i.e., smaller), signals from naturally occurring flaws in the hardware) to avoid overestimating NDE process/inspector capability. As described in NASA Standard 5009B, fatigue cracks are used for metallic hardware, which are assumed conservative as compared to many types of naturally occurring cracks that may develop. If alternate NDE simulated flaw types are used, then they must be approved by the responsible Fracture Control Board. If NDE simulated flaws based on material removal methods (e.g., electrical discharge machining (EDM) notches) are used, then transfer function approaches may be necessary to adjust the signal response from the NDE simulated flaws to those that might be detected from naturally occurring flaws of similar sizes. Additionally, corrections to signal responses might be required to adjust for additional factors that are non-representative of the flight hardware (e.g., material differences to include geometry, properties, and surface finish). Methods for such signal response adjustments are not included in this guidebook.

2.6.5 Sample Size Limitation

The potential consequences of using small sample sets in the LS-POD method should be recognized. The validation results shown in Appendix B illustrate the larger variability in the estimated value of the signal level that provides 90/95 POD as the sample size is decreased. This can lead to higher likelihood of failing the demonstration if a POF requirement cannot be met, and a larger non-conservative POD estimate in 5 percent of the cases when the estimate exceeds the 90-percent POD. If the LS-POD method is used for flawed sample sizes of less than 10, then additional review and approval should be obtained by the responsible Fracture Control Board and statistical expertise should be consulted. This review should include an evaluation of the degree of separation between the estimated POD and POF limits, the signal threshold level, and the likelihood of failure of the component. For application of the method with sample sizes of 10 or larger, the method can be used without additional review, but the results interpretation as defined in Section 2.5 must be followed.

3.0 NDE Field Application Process Monitoring

NDE process monitoring should be conducted to ensure that the inspector is providing the same or better capability as established during the demonstration testing. It is not feasible to represent every condition or source of variability in a POD demonstration test that might be observed in the inspection of flight hardware. As a result, it has been observed that the NDE inspector in practice does not provide the same capability that was demonstrated during testing. This can have disastrous effects if a flaw is missed that results in a catastrophic failure or have significant schedule and economic impacts if parts without flaws are erroneously rejected.

Basically, there are two sources of data that can be monitored during field application. The first source is noise measurements from regions assumed to have no flaws on the flight hardware. Monitoring noise data over time with respect to that established during the demonstration testing can provide indications of changes in the NDE instrumentation/measurement system, in the parts being inspected (e.g., material properties, surface conditions, geometry, etc.), and the performance of the inspector, which can impact the capability of the inspection. The second source of data is signal (and noise) responses from NDE standards (e.g., calibration blocks), which contain documented NDE simulated flaws. These data are used to standardize the instrument response and monitoring it over time can provide indications that the NDE instrument/measurement system is changing, which could negatively impact the capability of the NDE method/inspector.

There may be additional measurements that result from indications during the flight hardware inspections. However, these are not deemed to be as reliable for monitoring the NDE process as it may not be possible to independently confirm the presence of a flaw, nor characterize its size. Additionally, even if a flaw is independently confirmed, it is unlikely to be of the same size as used in the POD demonstration testing, so it is difficult to relate that signal response to the flaw signal distribution established during the POD demonstration testing.

3.1 Noise Monitoring

Assuming that the majority of components being inspected will not have flaws, recording signal levels from all field measurement opportunities can expand the unflawed signal database. This provides two opportunities to monitor and improve the inspection process.

Additionally, monitoring the noise distribution and its parameter estimates over time relative to the original LS-POD established noise distribution can provide insight into a variety of changes in the inspection process and the components being inspected. Such changes might negatively impact the capability of the inspection regarding the POF such to invalidate the inspection.

An unflawed measurement dataset with p measurements is collected in the field as described in Section 2.1, where $p \geq 40$, and the following parameters are estimated in same manner as the demonstration experiment.

The mean signal in the absence of flaws ($\bar{y}_{N_{\text{field}}}$) is calculated by:

$$\bar{y}_{N_{\text{field}}} = \frac{\sum_{i=1}^{p} y_{N_{\text{field}_i}}}{p}$$

The standard deviation of noise, or signals in the absence of flaws ($s_{N_{\text{field}}}$) is calculated by:

$$s_{N_{\text{field}}} = \sqrt{\sum_{i=1}^{p} \frac{(y_{N_{\text{field}_i}} - \bar{y}_{N_{\text{field}}})^2}{p-1}}$$

Specifically, the signal level corresponding to a 1/95 POF for flight hardware measurements ($y_{1/95\ \text{POF}_{\text{field}}}$) can be defined as:

$$y_{1/95\ \text{POF}_{\text{field}}} = \bar{y}_{N_{\text{field}}} + k_{1_{N_{\text{field}}}} * s_{N_{\text{field}}}$$

From Appendix A, it is shown that the expected variability in the 1/95 POF limit from the LS-POD testing is $0.10(y_{1/95\ \text{POF}} - \bar{y}_N)$ for the noise sample size of 40. Thus if:

$$y_{1/95\ \text{POF}_{\text{field}}} > y_{1/95\ \text{POF}} + 0.10(y_{1/95\ \text{POF}} - \bar{y}_N)$$

then the noise has increased beyond expected variability, and the NDE process should be evaluated to identify the possible cause(s). Additionally, if

$$y_{1/95\ \text{POF}_{\text{field}}} > y_{dec}$$

then the POF requirement is not met in the flight hardware inspections, and the inspection process should be stopped to determine the cause(s), and corrections made and verified before valid inspections are continued. The approved inspection procedure developed and approved by the responsible NDE engineer should specify the intervals and the numbers of noise measurements that should be acquired and periodically analyzed to provide this NDE process monitoring. The procedure should specify the actions to be taken should the noise process monitoring threshold be violated.

With repeated inspections, an expanded unflawed signal database that combines all unflawed field measurement data can be constructed, and it will provide an improved estimate of the noise distribution assuming that the noise from field measurements is consistent with that developed during the demonstration testing. The resulting improved estimates of the noise signal distribution parameters may result in a lower signal level required to meet the POF requirement, providing room for potentially lowering the decision threshold to provide a higher POD. Any such changes to the POD testing established noise distribution parameters and resulting decision threshold should be documented as an addendum in the POD demonstration test report and in the NDE procedure.

Instrument Standardization as defined in NASA Standard 5009B is the adjustment of an NDE instrument response using an appropriate reference standard with known size discontinuities such as EDM slots and flat bottom holes, to obtain or establish a known and reproducible response. Instrument standardization is required in addition to instrument calibration. The NDE procedure should define when instrument standardization is to be performed, but it is typically required at a minimum before and after an NDE inspection. In addition, it may be required at certain intervals (i.e., periods of time or number of parts inspected) for inspections that occur over long time periods. NASA 5009 recommends that instrument standardization should be carried out using a minimum of three data points demonstrating expected correlation between signal response and discontinuity size. While there are differences in standardization dependent on the method or the particular inspection application, the following general recommendations should be considered and incorporated as appropriate into the inspection procedure. One of the NDE standards comprising the three-point instrumentation standardization procedure should be designed to produce a signal consistent with that produced by the desired target flaw size. The other NDE standards can be designed to produce signals consistent with larger flaw sizes, but those larger flaw sizes should be in the range expected to form in the part and be of interest for detection. The degree of repeatability of the measured signal responses from the NDE standards should be documented in the procedure, with established limits on how much instrument adjustment (i.e., gain/attenuation and offset) is permitted to achieve the desired signal response levels. In addition to monitoring the signal levels from the NDE standards, noise signals should be measured on defined regions of the standards where no flaw exists. Minimum acceptable NDE standard signal/noise (S/N) levels should be established in the NDE method development and documented in the procedure. The S/N should be evaluated at each instrument standardization opportunity to ensure that it exceeds this minimum specified value.

There are additional signal and noise quality metrics that can be used to monitor the NDE process. These are described by Koshti (2021) and can be evaluated for applicability for the particular inspection method and application. If additional metrics are used, they should be described, and detailed instructions provided in the NDE procedure.

4.0 References

Koshti, A.: "Using requirements on merit ratios for assessing reliability of NDE flaw detection," *SPIE Smart Structures and NDE*, March 2021, Long Beach, CA, Abstract accepted, 2021.

MIL-HDBK-1823A: "Nondestructive Evaluation System Reliability Assessment," 2009.

NASA Standard 5009B: "Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components," 2019.

Olin, B. D.; and Meeker, W. Q.: "Applications of Statistical Methods to Nondestructive Evaluation," *American Statistical Association and the American Society for Quality Control*, *Technometrics*, 38(2), 95-112, 1996.

Rummel, W.: "Recommended Practice for a Demonstration of Nondestructive Evaluation (NDE) Reliability on Aircraft Production Parts," *Materials Evaluation*, 40(9), 922-932, 1982.

Swets, J. A.: "Assessment of NDT Systems – Part I: The Relationship of True and False Detections," *Materials Evaluation*, 41, 1294-1298, 1983a.

Swets, J. A.: "Assessment of NDT Systems – Part II: Indices of Performance," *Materials Evaluation*, 41, 1300-1303, 1983b.

Walpole, R. E.; Myers, R. H.; and Myers, S. L.: *Probability and Statistics for Engineers and Scientists*, 6th edition, Prentice Hall, Chapter 10, 1998.

Appendix A. Statistical Details of the Limited Sample Probability of Detection Method

Overview

Nondestructive evaluation (NDE) techniques are used to make decisions on whether or not a flaw indication is present in a test specimen or hardware. In practice, a measurement from an NDE technique with a continuous signal is compared to a decision threshold, where signals that exceed the decision threshold indicate the presence of a flaw. NDE probability of detection (POD) studies are conducted to establish the detectable flaw size associated with specified inferential risks in declaring a flaw to be present or not. Limited Sample POD (LS-POD) is a type of a POD study that restricts its attention to a single, target flaw size with a minimal sample size. Table A.1 is adapted from classical statistical hypothesis testing to illustrate the potential outcomes from an NDE inspection decision.

Table A.1. Potential Outcomes of an NDE Inspection Highlighting Inferential Risks

		Known State of Whether a Flaw is Present	
		No Flaw is Present	**Flaw is Present**
NDE Decisions	**Declare no Flaw is Present**	Correct Decision (True Negative)	Incorrect Decision (False Negative)
	Declare a Flaw is Present	Incorrect Decision (False Positive) **POF**	Correct Decision (True Positive) **POD**

LS-POD focusses on Probability of a False Call (POF) and POD, and the cases of a true negative or false negative are not considered. Differing from traditional statistical hypothesis testing, LS-POD seeks to mimic established NDE practice as described in MIL-HDBK-1823A (2009) in considering POF and POD independently. However, POD and POF are implicitly related in the LS-POD guidance and interpretation of the results in Section 2.5.

The LS-POD method provides an estimated POF limit on the NDE signal below which there is no detected indication that is considered distinguishable from inherent measurement variability (i.e., noise), and this method provides an estimated POD limit on the NDE signal above which a target flaw size is considered to be reliably detected. Based on the estimated 1/95 POF and 90/95 POD limits, the feasibility of a decision threshold can be evaluated for flight hardware inspections.

Comment on LS-POD Notation

LS-POD implements notation that is consistent with MIL-HDBK-1823A and standard statistical nomenclature, while not propagating previously cited notational deviations contained in MIL-HDBK-1823A where continuous-response NDE techniques are commonly referred to as $\hat{a}$-vs-a data. The use of $\hat{a}$ is problematic since it implies a predicted value of a in standard statistical convention, when in fact $\hat{a}$ denotes the measured inspection signal and not an estimated quantity. This potential confusion has been cited in the literature, for example, Olin and Meeker (1996) states, "When the output from an NDE measurement process is a continuous response that is used to estimate flaw size a, the response is often denoted, in the NDE literature, by $\hat{a}$, (not to be confused with a parameter estimate or estimated mean response as is the convention in statistics)." In the NDE literature, authors often generalize their notation to using y for the signals and x for the flaw size in the presentation of their methodology, see Section G.3 in MIL-HDBK-1823A. This is the notational approach adopted by LS-POD.

Data Acquisition and Descriptive Statistics

A random sample of NDE flaw signal measurements, y_F, are acquired from inspections on m specimens representative of flight hardware (i.e., similar material, form, and surface finish) containing known NDE simulated flaws of a single flaw size. In addition, a random sample of NDE unflawed signals, y_N, are acquired from inspections performed on n unflawed specimens or inspections in regions of the flawed specimens where no flaws are known to exist. The flaw and unflawed measurements are considered to be independent and identically distributed, from their own respective distributions, referred to as the flaw and unflawed distributions.

For the statistical analyses that follow, the test specimens are assumed representative of the flight hardware and no transfer function is applied.

For the flaw signal measurements, the mean signal in the presence of target flaws ($\bar{y}_F$) is calculated by:

$$\bar{y}_F = \frac{\sum_{i=1}^{m} y_{F_i}}{m}$$

The standard deviation of signals in the presence of target flaws (s_F) is calculated by:

$$s_F = \sqrt{\sum_{i=1}^{m} \frac{\left(y_{F_i} - \bar{y}_F\right)^2}{m-1}}$$

Since LS-POD's objective is to utilize a small sample size, it is impractical to reliably identify the underlying distributional model and therefore a Normal distribution is assumed. If the sample size is large enough, the assumption of a Normal distribution could be evaluated with a quantile-quantile plot as described in Appendix G.3.5.3 of MIL-HDBK-1823A.

For the unflawed signal measurements, the mean signal in the absence of flaws ($\bar{y}_N$) is calculated by:

$$\bar{y}_N = \frac{\sum_{i=1}^{n} y_{N_i}}{n}$$

The standard deviation of noise, or signals in the absence of flaws (s_N) is calculated by:

$$s_N = \sqrt{\sum_{i=1}^{n} \frac{(y_{N_i} - \bar{y}_N)^2}{n-1}}$$

A Normal distribution is assumed for the unflawed signals or an appropriate transformation of the unflawed signals. For example, a logarithm transformation might be considered to avoid negative signal values for some NDE techniques. With a minimum sample size of 40 unflawed signal measurements, it is practical to evaluate the fit of the Normal distribution with a quantile-quantile plot, as illustrated in Appendix G.3.5.3 of MIL-HDBK-1823A. Evaluating the Normal distribution's quality of fit is not an attempt to prove that the data are normally distributed, rather it is a check for gross violations of that assumption.

Establishing Signal Limits for a Specified POF and POD

Statistical tolerance bounds for a normal distribution are employed to estimate the signal limits for a specified POF and POD. A tolerance interval covers a specified proportion of a distribution with a stated confidence, see Meeker et al. (2017) for additional details.

For POD, a signal limit that will be exceeded by at least 90 percent of the population of flaw signal measurements from the target flaw size with 95-percent confidence is known as a 90/95 limit. In particular, it is one side of the tolerance interval, known as a confidence bound, since only the signals that exceed the limit are of interest. Assuming that the NDE technique signal level increases with flaw size and is sufficiently separate from saturation or censoring, LS-POD is conceptually similar to the approach that MIL-HDBK-1823A uses to estimate the 90/95 detectable flaw size.

Since the estimated signal limit is based on a sample of the population, the analysis includes a factor, k_1, which compensates for the estimation uncertainty associated with the sample size. For small sample sizes, there is a larger uncertainty in estimating the POD limit.

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

$$\delta = z_p * \sqrt{m}$$

$$k_{1_F} = \frac{t_{(\alpha, m-1, \delta)}}{\sqrt{m}}$$

where z_p is the critical value from a standard normal distribution for $p = 0.90$ based on a 90-percent proportion of the population, $\alpha = 0.95$ for a 95-percent confidence level, and δ is the non-centrality parameter. Appendix C provides a table of 90/95 k_{1_F} factors as a function of sample size. The signal limit, which provides 90-percent POD at 95-percent confidence, is estimated by:

$$y_{90/95\ \mathrm{POD}} = \bar{y}_F - k_{1_F} * s_F$$

Following a similar approach for POF, a signal limit that exceeds 99 percent of the population of unflawed signal measurements with 95-percent confidence is estimated. For clarity in the NDE vernacular, it is called the 1/95 limit rather than the 99/95 limit. The k_{1_N} value is estimated as follows.

$$\delta = z_p * \sqrt{n}$$

$$k_{1_N} = \frac{t_{(\alpha, n-1, \delta)}}{\sqrt{n}}$$

where z_p is the critical value from a standard normal distribution for $p = 0.99$ based on a 99-percent proportion of the population, $\alpha = 0.95$ for a 95-percent confidence level, and δ is the non-centrality parameter. Appendix D provides a table of 1/95 k_{1_N} factors as a function of sample size. The signal limit which will provide 1-percent POF at 95-percent confidence is estimated by:

$$y_{1/95\ \mathrm{POF}} = \bar{y}_N + k_{1_N} * s_N$$

Since NASA Standard 5009B does not require a specific POF value, it is expected that application specific decisions could be made to accept a lower or higher POF. As an example of k_{1_N} values for other POF values, Table A.2 provides associated k_{1_N} factors for an unflawed sample size of 40.

Table A.2. Unflawed k_1 Factor Values for Sample Size of $n = 40$ for Different POF Probabilities.

	Probability of False Calls (POF) with 95 percent Confidence for $n = 40$				
	0.1%	**1%**	2%	5%	10%
k_{1_N}	3.865	**2.941**	2.613	2.125	1.697

The feasibility of a LS-POD demonstration for a target flaw size can be evaluated based on the relationship between these estimated POF and POD limits, and their relationship to a decision threshold as discussed in Section 2.5.

Graphical Illustration of the Limited Sample POD Method

In Figure A.1, simulated unflawed and flaw signal measurements are illustrated by their respective histograms. The unflawed and flawed normal distributions fit to the data are overlayed, along with the POF and POD limits and a decision threshold. In this illustration of a successful LS-POD demonstration, the POF limit falls below the decision threshold and the POD limit is above the decision threshold.

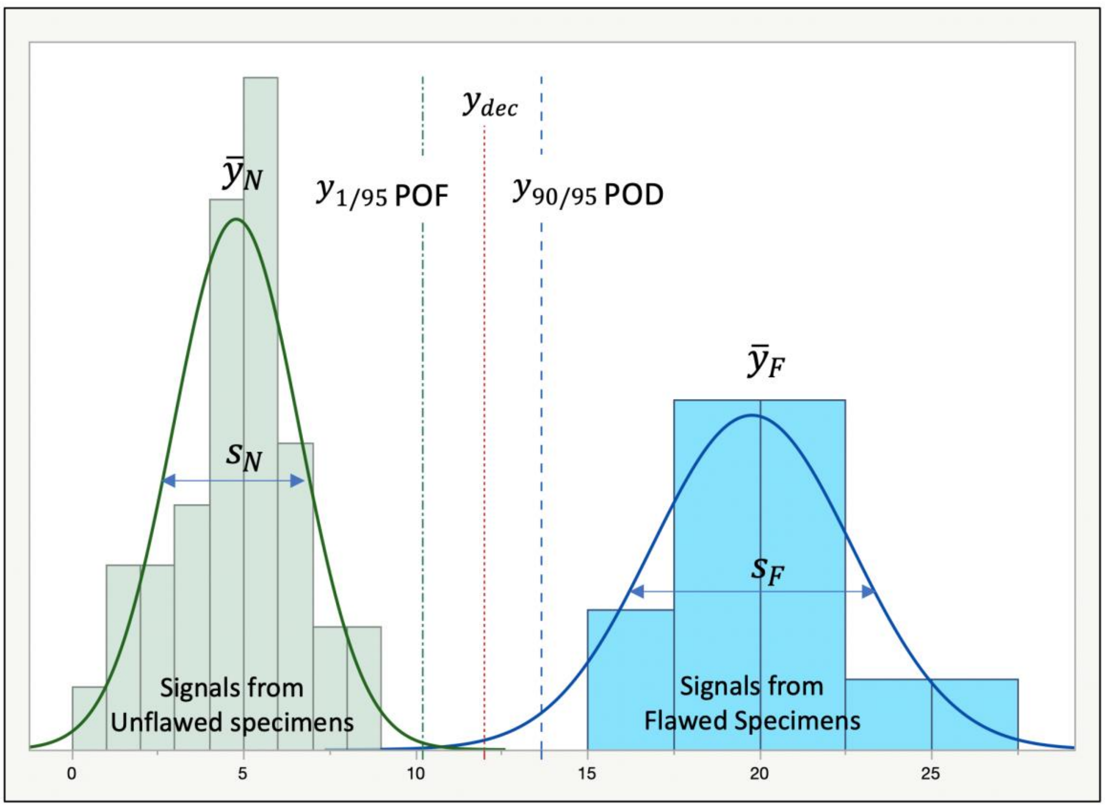

Figure A.1. Graphical illustration of LS-POD method results.

Process Monitoring Statistical Details

In Section 3.0, process monitoring is recommended by comparing field-based POF limit, $y_{1/95\ \mathrm{POF_{field}}}$, to the demonstration POF limit, $y_{1/95\ \mathrm{POF}}$. There is variability in the estimated $y_{1/95\ \mathrm{POF}}$, therefore sufficient separation is required when making the comparison to protect against spurious decision outcomes. To illustrate the need for this separation, if a noise measurement experiment was conducted twice with the exact same specimens, there would likely be a numerical difference in the estimated $y_{1/95\ \mathrm{POF}}$ limit when there is no cause for alarm that something has changed, since they are the same specimens.

A practical approach to accommodate this expected variability in comparing $y_{1/95\ \mathrm{POF}}$ limits is proposed by recognizing that a tolerance interval is equivalent to a confidence interval around a proportion (the 0.99 percentile) of the population. In Figure A.2, an unflawed measurement distribution is illustrated with a superimposed distribution around the upper 1st percentile that represents the precision of the percentile estimate given the sample size and the variability of the unflawed measurements in the samples. Figure A.2 shows a third, smaller superimposed distribution of the estimated 1/95 POF limit ($y_{1/95\ \mathrm{POF}}$).

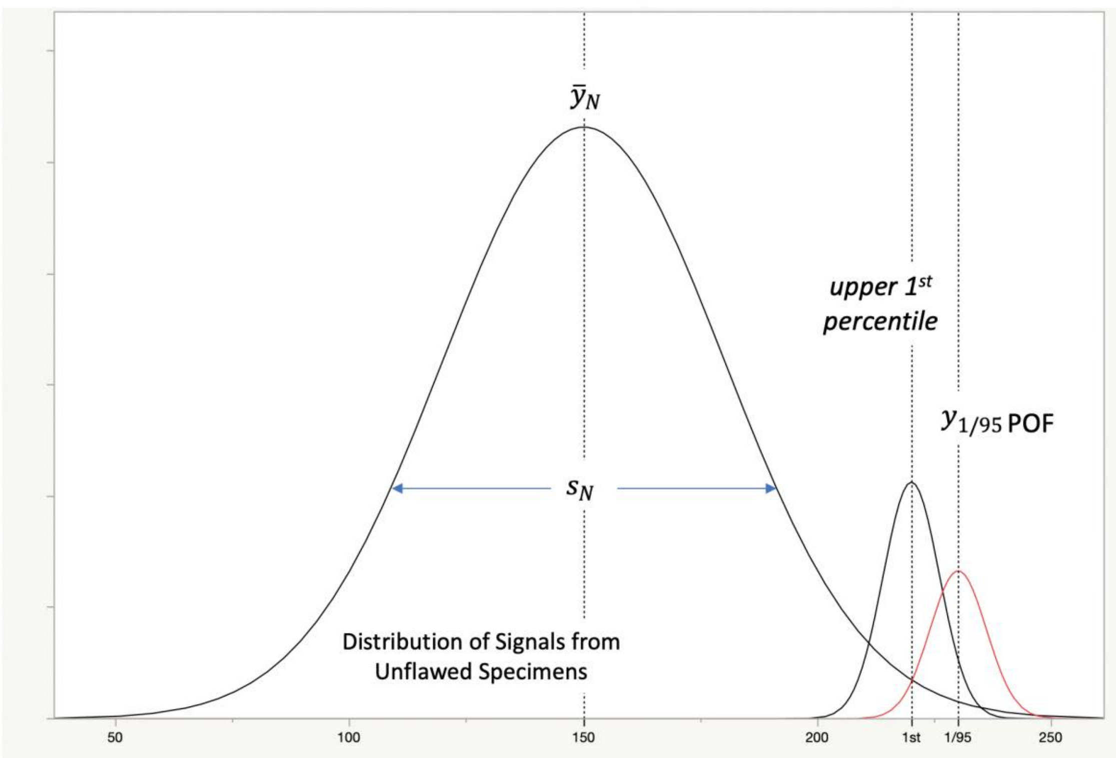

Figure A.2. Illustration of the Unflawed measurement distribution, variability of the estimated quantile, and variability of the $y_{1/95\ \mathrm{POF}}$ limit.

The standard deviation of the estimated $y_{1/95\ \mathrm{POF}}$ is proportional to the variability of the unflawed measurements, s_N, and the sample size, n. It can be shown that the standard deviation of $y_{1/95\ \mathrm{POF}}$ can be approximated by a percentage of the POF limit minus the mean of the unflawed distribution, $(y_{1/95\ \mathrm{POF}} - \bar{y}_N)$. This expression is algebraically equivalent to $k_{1_N} * s_N$, and since k_{1_N} is a function of n it is clear that the variability of the 1/95 POF signal level is a function of sample size.

However, using $(y_{1/95\ \mathrm{POF}} - \bar{y}_N)$ in practice is considered advantageous, since those values are more readily accessible from the LS-POD demonstration results. Table A.3 provides the approximate percentages for three sample sizes.

Table A.3. Standard Deviation of $y_{1/95\ \mathrm{POF}}$ as a Function of Sample Size

Sample Size (n)	Standard Deviation of $y_{1/95\ \mathrm{POF}}$ Limit as a Percentage of $(y_{1/95\ \mathrm{POF}} - \bar{y}_N)$
40	**11%**
60	9%
100	7%

It can be shown through an analytical derivation that the percentage values in the table are a function of sample size, n, only, which makes them generally applicable regardless of the estimated sample mean and standard deviation.

The LS-POD method uses $0.10(y_{1/95\ \mathrm{POF}} - \bar{y}_N)$ as the minimum acceptable difference, as specified in Section 3.1, starting with paragraph 2.

$$y_{1/95\ \mathrm{POF}_{\mathrm{field}}} > y_{1/95\ \mathrm{POF}} + 0.10(y_{1/95\ \mathrm{POF}} - \bar{y}_N)$$

When making a comparison to a field estimate of the limit, $y_{1/95\ \mathrm{POF}_{\mathrm{field}}}$, this margin on the difference provides a measure of protection against potentially spurious indications of increases in the 1/95 POF limit.

This guidance on the variability of the $y_{1/95\ \mathrm{POF}}$ limit assumes replicated experiments from the same set of specimens, and it does not account for additional variability in comparing LS-POD specimens to field specimens. In current NDE practice, conducting a dedicated unflawed specimen measurement experiment is not typically performed. While multiple unflawed measurement experiments could be conducted to assess the demonstration-to-hardware inspection variability, until LS-POD is implemented and evaluated in practice, it was considered not practical to require additional unflawed measurements experiments to estimate this additional variability. Rather, the proposed factor of 0.10 is considered conservative for sample sizes exceeding the minimum specified of 40, and the risk of some investigations spuriously detecting differences between $y_{1/95\ \mathrm{POF}}$ and $y_{1/95\ \mathrm{POF}_{\mathrm{field}}}$ was considered acceptable.

References:

Meeker, W. Q.; Hahn, G. J.; and Escobar, L. A.: *Statistical Intervals: A Guide for Practitioners and Researchers*, 2nd Edition, John Wiley & Sons Inc., 2017.

Appendix B. Demonstration and Validation of the Limited Sample POD Method

Two datasets are used to demonstrate the application of the Limited Sample POD (LS-POD) method in this Appendix. The data required for LS-POD are flaw signal measurements, y_F, acquired from inspections of specimens containing known NDE simulated flaws of a single target flaw size, and unflawed signals, y_N, acquired from inspections performed on unflawed specimens. A decision threshold may be specified prior to conducting a LS-POD, or it may be chosen based on the LS-POD analysis results.

Demonstration Dataset 1 – Example 1 from MIL-HDBK-1823A

This dataset comes from Section G.3.4 in MIL-HDBK-1823A, and it is included in the companion R-code distribution package as "EXAMPLE 1 a.hat vs a.xls." There are 92 measurements of varying flaw sizes are provided in Table B.1, and there are no censored observations. MIL-HDBK-1823A recommends modeling the signal response as a function of the transformed flaw size as $\log(a)$. The raw data from Table B.1 are plotted using a log scale for flaw size in Figure B.1.

Table B.1. Example 1 dataset used in MIL-HDBK-1823A Appendix G.

Flaw Size	Signal	Flaw Size	Signal	Flaw Size	Signal	Flaw Size	Signal
6	192	13	139	17	461	29	777
6	166	13	342	17	483	29	1040
7	137	13	614	18	599	30	753
7	97	13	537	18	507	30	778
7	147	13	370	18	728	31	729
8	221	14	439	19	344	32	811
8	148	14	138	19	514	32	1208
8	152	14	50	19	568	32	864
9	376	14	621	19	843	32	938
9	568	14	610	20	645	33	865
9	324	15	509	20	838	33	853
9	363	15	482	20	563	34	740
10	87	15	353	20	638	35	830
10	198	15	260	21	751	39	1159
10	394	15	381	22	736	41	743
10	413	15	415	23	703	42	1176
11	290	15	548	24	756	45	1079
11	318	16	537	24	731	48	1344
12	50	16	672	25	713	50	1277
12	50	17	352	27	787	50	1335
12	304	17	671	27	763	52	1283
12	414	17	437	27	963	53	1375
13	519	17	435	28	818	65	1189

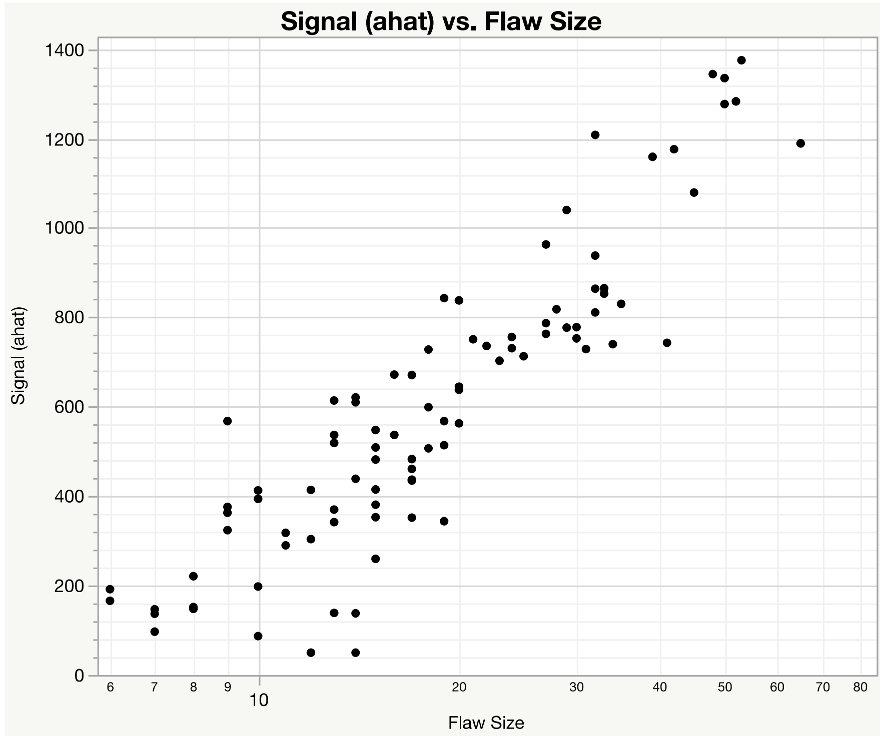

Figure B.1. Signal versus flaw size from Example 1 in MIL-HDBK-1823A.

Estimating Probability of Detection (POD) Limit, $y_{90/95\ \mathrm{POD}}$

For a target flaw size of 17, flaw signals associated with flaw sizes of 16, 17, and 18 were chosen to simulate a LS-POD study, providing 11 measurements shown in Table B.2. Treating this range of flaw sizes as coming from a collection of specimens of a single target flaw size mimics practice where variability in the nominal flaw size is anticipated.

Table B.2. Flaw Signal Measurements from Example 1 in MIL-HDBK-1823A

Flaw Size	Signal
16	537
16	672
17	352
17	671
17	437
17	435
17	461
17	483
18	599
18	507
18	728

The sample mean and standard deviation are estimated according to Appendix A as $\bar{y}_F = 534.7$ and $s_F = 118.6$. From Appendix C, for sample size of 11, $k_{1_F} = 2.275$, and $y_{90/95\ \mathrm{POD}}$ is estimated as:

$$y_{90/95\ \mathrm{POD}} = 534.7 - 2.275 * 118.6 = 265$$

Therefore, this represents the maximum decision threshold that maintains 90-percent POD with 95-percent confidence. Signals above $y_{90/95\ \mathrm{POD}} = 265$ would be called as a hit in an NDE inspection.

Estimating Probability of False Call (POF) Limit, $y_{1/95\ \mathrm{POF}}$

A dedicated experiment to measure unflawed specimens was not conducted in the dataset provided in MIL-HDBK-1823A; however, the Handbook suggests that signals from flaw sizes below 8.5 are unrelated to flaw size, and therefore they are considered to be unflawed noise measurements. Therefore, the unflawed measurement data include 8 signal measurements associated with flaw sizes of 6, 7, and 8 with an estimated sample mean of 157.5 and a standard deviation of 37.0. To generate a minimum specified sample size to demonstrate LS-POD, 40 unflawed measurements were simulated from a normal distribution based on the sample mean and standard deviation. Table B.3 provides these 40 simulated unflawed measurements.

Table B.3. Simulated Unflawed Noise Measurements for Example 1 from MIL-HDBK-1823A

121	157	138	156	171	208	218	160	147	192
149	116	156	182	128	119	182	81	151	123
102	164	166	199	154	221	152	133	185	149
126	111	165	129	170	203	177	223	153	137

A histogram of the simulated unflawed measurements overlaid with an estimated normal distribution is shown in Figure B.2.

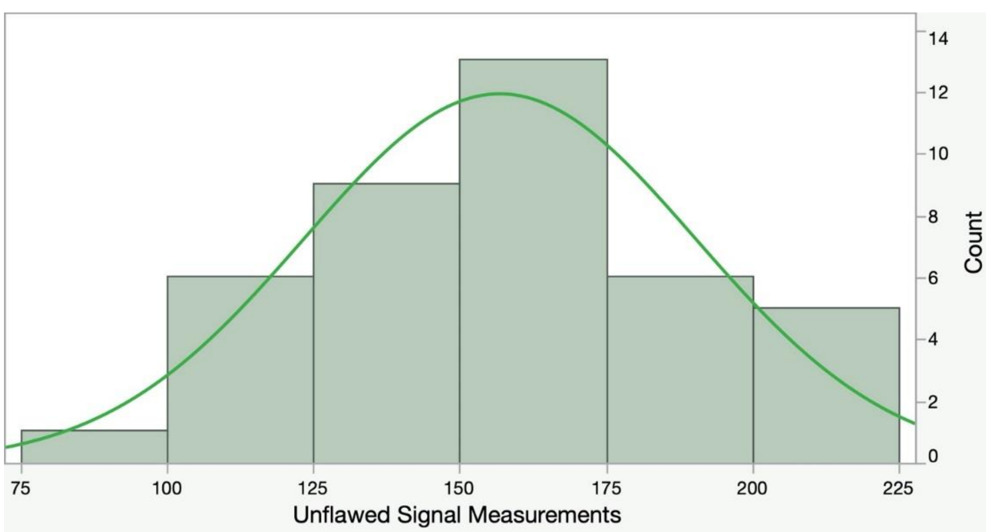

Figure B.2. Unflawed noise measurements from Example 1 in MIL-HDBK-1823A.

The estimated sample mean and standard deviation of the unflawed measurements are $\bar{y}_N = 156.8$ and $s_N = 33.5$. From Appendix D, for a sample size of 40, $k_{1_N} = 2.941$, and $y_{1/95\ \mathrm{POF}}$ is estimated as

$$y_{1/95\ \mathrm{POF}} = 156.8 + 2.941 * 33.5 = 255$$

Therefore, signals below $y_{1/95\ \mathrm{POF}} = 255$ would not be called as a hit by the NDE inspection to maintain a POF of 1 percent with 95-percent confidence.

Interpretation of the LS-POD Results

Based on the guidance provided in Section 2.5, the result that $y_{90/95\ \mathrm{POD}} \geq y_{1/95\ \mathrm{POF}}$ satisfies the first criteria of a successful demonstration at 90/95 POD and 1/95 POF. A decision threshold value between the POF and POD limits, where $255 \leq y_{dec} \leq 265$, is considered acceptable for the inspection of the flight hardware.

In this particular example, the 1/95 POF limit is close to the 90/95 POD limit, which indicates that the unflawed and flawed distributions are relatively close together. In cases like this, a LS-POD demonstration may not be initially successful, and it may require additional flawed specimens, acceptance of an increased POF, or a larger target flaw size.

Simulation Study to Illustrate the Effect of Sample Size on the Estimated POD Signal Limit

A simulation study was conducted to illustrate the performance characteristics of LS-POD method using the Example 1 dataset. Replicated LS-POD demonstration tests were simulated with sample sizes ranging from 10 to 30, and the $y_{90/95\ \mathrm{POD}}$ POD signal limit is estimated for each simulated demonstration test.

For the flaw measurement distribution, it was assumed that the flaw signal population is normally distributed with a mean and standard deviation of 535 and 119, respectively, which are based on the estimated sample mean and standard deviation in the previous example. Since these are the population parameters (unknown in practice), the true 90-percent POD limit can be found directly from the 10th percentile of the normal distribution as 382. Comparing the estimated POD limits to the true POD limit of 382, it is expected that in about 5 percent of the cases the estimated $y_{90/95\ \mathrm{POD}}$ will exceed 382. When the $y_{90/95\ \mathrm{POD}}$ POD signal limit exceeds the true 90-percent POD limit, the method is non-conservative. Minor variation in the percentage that exceeds the true POD limit is a function of the number of simulation cases performed.

To consider the relationship of the POD and POF signal limit simultaneously, it is assumed that the unflawed signal population is normally distributed with a mean and standard deviation of 157 and 34, respectively, which are based on the estimated sample mean and standard deviation of the unflawed measurements in previous example. In this simulation study, the effect of varying the unflawed measurement sample size on the 1/95 noise limit is not considered. Since these are population parameters, the true 1-percent (upper 99-percent) limit is found directly from the 99th percentile of the normal distribution as 236.

In Figure B.3, the results from 500 simulated LS-POD demonstration tests with samples sizes of 10, 15, 20, 25, 30 are shown along with reference lines at the true POD limit and true POF limit based on the simulation parameters.

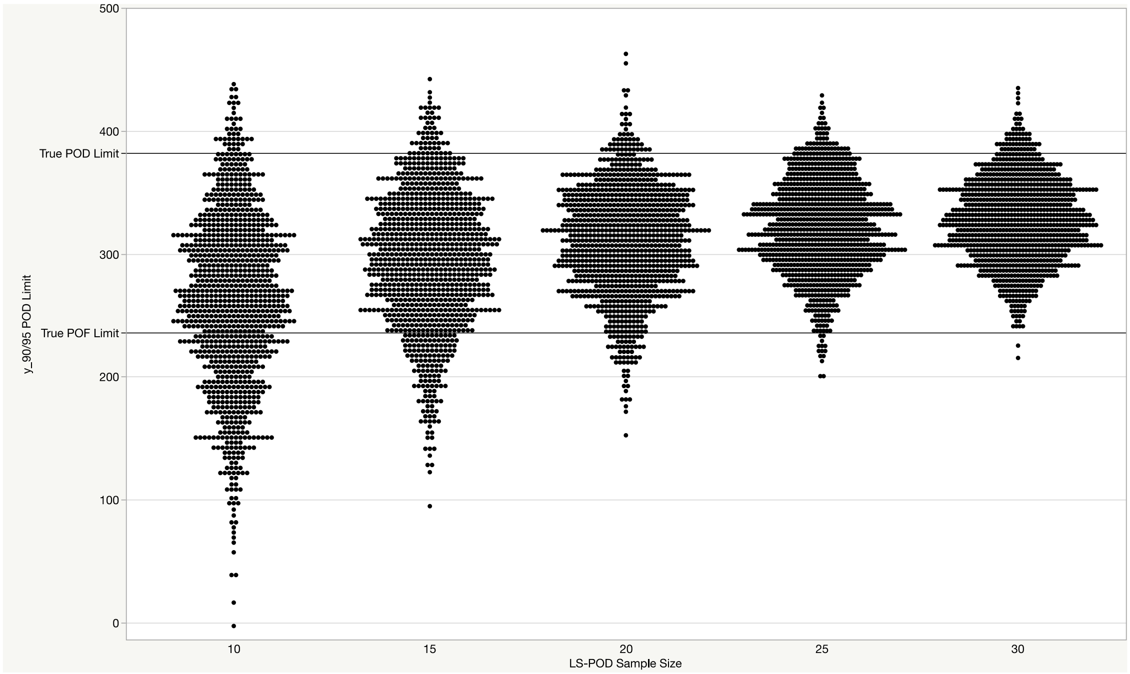

Figure B.3. Simulated POD limits as a function of sample size based on Example 1 from MIL-HDBK-1823A.

Discussion of the Simulation Study Results

Consistent with theoretical expectations, Figure B.3 shows that the variability in 90/95 POD limits due to estimation uncertainty reduces with an increasing number of flawed specimens. The figure illustrates the 5-percent risk of estimating non-conservative POD limits that are larger than the true 90-percent POD limit, and how that risk is consistent across the sample sizes. However, the frequency of estimating a POD limit below the true POF requirement is higher for smaller sample sizes. As an example, for a sample size of 10 about 30 percent of the estimated POD limits fall below the POF limit in this simulation case. At a sample size of 20, the POD limit is below the POF limit in about 5 percent of the cases. Finally, it can be observed that for smaller sample sizes when the 90/95 POD limit exceeds the true POD, it may exceed that value by a larger magnitude due to the variability associated with the estimates.

Tying this simulation study back to the LS-POD analysis of the Example 1 dataset, the estimated POD and POF limits are relatively close together at 265 and 255, respectively, which indicates that unflawed and flaw signal distributions are not well separated. In general, to meet the POD requirement with acceptable POF, the sample size is a function of the separation between the POD and POF limits. When the POD and POF limits are close to each other as discussed in Section 2.5, increasing the sample size, increasing the target flaw size, or allowing a higher POF may be required. Smaller sample sizes may be acceptable if the POD and POF limits are appropriately separated (See Section 2.5).

A Numerical Comparison to MIL-HDBK-1823A POD Analysis

To illustrate the relationship between LS-POD and a standard $\hat{a}\ vs\ a$ POD study described in Appendix G of MIL-HDBK-1823A is performed on Example 1 dataset. In MIL-HDBK-1823A's analysis of this dataset, the decision threshold is specified at $y_{dec} = 200$, resulting in an estimated

$a_{90/95}$ flaw size of 13.7; however there is a POF of about 11 percent. This POF was considered too high to demonstrate LS-POD. Therefore, a decision threshold of $y_{dec} = 265$ was chosen since it is the minimum signal limit for 90/95 POD of the target flaw size based on the LS-POD method, and it exceeds the POF limit of 255.

The data were modeled with flaw sizes above 8.5 to be consistent with MIL-HDBK-1823A's assertion that signals below a flaw size of 8.5 are unrelated to flaw size. Therefore, $m = 92 - 8 = 84$ measurements are utilized in the modeling of $\hat{a}\ vs\ \log(a)$.

Figure B.4 shows the $\hat{a}\ vs\ \log(a)$ measurements, an estimated linear model, and the 90-percent two-sided prediction interval, which is equivalent to a 95-percent one-sided prediction bound.

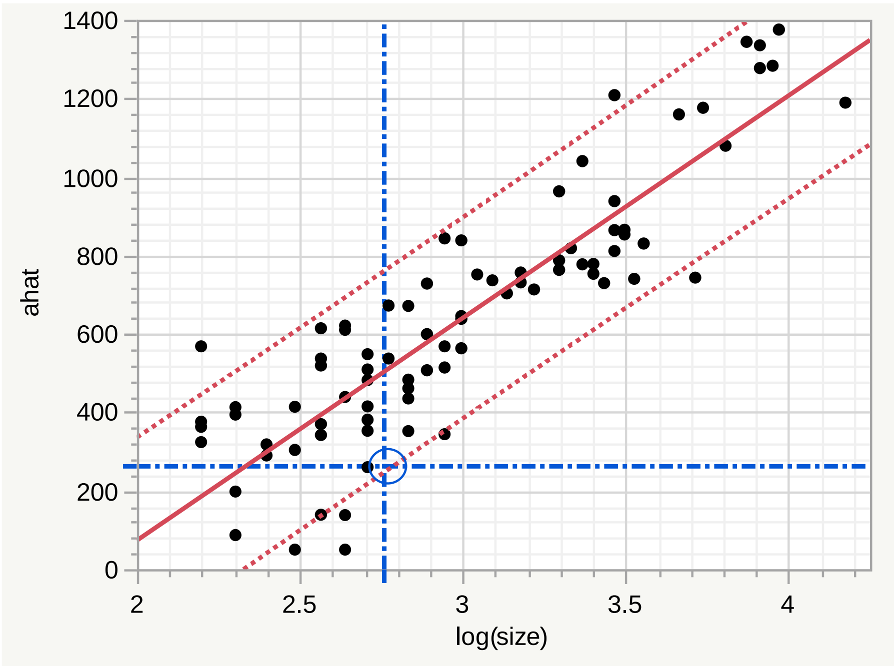

Figure B.4. $ahat\ vs\ \log(a)$, estimated linear model, 90-percent two-sided prediction interval, with guidelines at the decision threshold and $a_{90/95}$ flaw size.

Using the notation from MIL-HDBK-1823A, the coefficients of the linear model are $\hat{\beta}_0 = -1060.00$ and $\hat{\beta}_1 = 566.74$. The residual error from the regression model is $\hat{\tau} = 153.13$. It can be shown that the $a_{90/95}$ flaw size is estimated by

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

where y_{dec} is the specified decision threshold. Using the formulation provided in Appendix A, the 90/95 k_1 value is 1.556, which is based on a sample size of 82 degrees of freedom associated with the residual error. The $a_{90/95}$ flaw size is estimated as

$$\log(a_{90/95}) = \frac{1}{566.74}(265 - (-1060.00) + 1.556 * 153.13) = 2.758$$

$$a_{90/95} = e^{2.758} = 15.8$$

In Figure B.4, a horizontal dotted guideline is shown at the signal decision threshold of 265, and a vertical dotted guideline is shown at 2.758 in the log (a) axis. Note that their intersection is near the lower prediction interval that represents a one-sided 95-percent signal limit at the decision threshold.

As an approximate comparison to LS-POD's results, the average flaw size from Table B.2 is 17.1. Therefore, using a decision threshold of 265 for LS-POD results in a larger detectable flaw size than a MIL-HDBK-1823A POD study. This result is expected due to the smaller sample size used in LS-POD resulting in a larger uncertainty reflected in the k_1 values of 2.275 and 1.556 for the LS-POD and MIL-HDBK-1823A POD estimates, respectively. As a cautionary note, LS-POD cannot be used to estimate the $a_{90/95}$ flaw size since it is derived from measurements of a single target flaw size.

Demonstration Dataset 2 – Eddy Current Inspection from Raised Head Fasteners

This dataset is simulated based on a published data set from eddy current inspections around raised head fasteners for fatigue cracks emanating from the fastener holes, see Forsyth et al. 2010.

The original data had censoring at the upper end of the signal range, and it was modeled using the MIL-HDBK-1823A guidance for fitting a linear fit to the relationship of signal to crack length on the surface. The estimated model parameters were used to simulate a new set of data without censoring that allows the estimation of signal distributions across a wider range of crack sizes to demonstrate the LS-POD method. The simulated inspection measurements are plotted in Figure B.5 and tabulated in Table B.4.

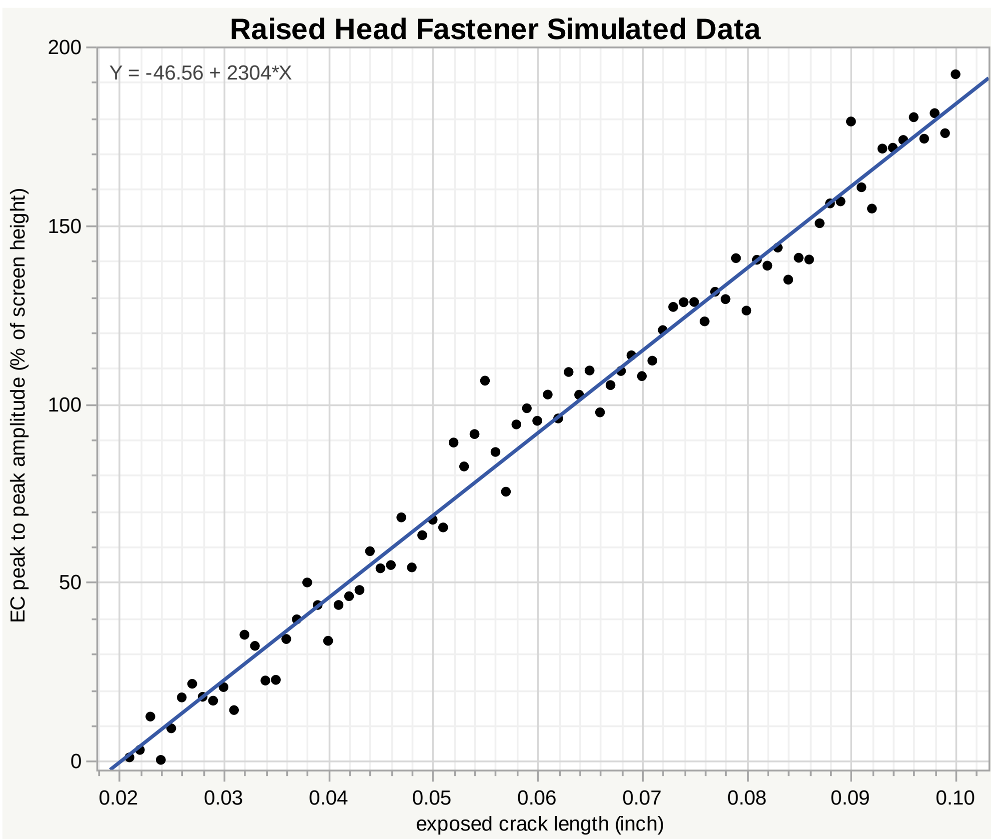

Figure B.5. Raised head fastener $ahat$ vs. a simulated data with linear model.

The following linear model is assumed.

$$\hat{a} = \beta_0 + \beta_1 a + \varepsilon$$

where $\hat{a}$ is the eddy current signal, a is the defect size, β_0 is the y intercept of the model, β_1 is the slope, and the residuals ε are normally distributed with a zero mean and standard deviation of τ. A regression is performed to estimate the parameters of the linear model based on the data in Table B.4 as $\hat{\beta}_0$ = -46.6, $\hat{\beta}_1$ = 2304, and $\hat{\tau}$ = 7.30.

Table B.4. Raised Head Fastener $ahat$ vs. a Simulated Data

defect size (inch)	signal (arbitrary)		defect size (inch)	signal (arbitrary)
0.021	1.01		0.061	102.56
0.022	3.12		0.062	95.87
0.023	12.44		0.063	108.86
0.024	0.32		0.064	102.49
0.025	9.14		0.065	109.28
0.026	17.80		0.066	97.58
0.027	21.61		0.067	105.20
0.028	17.97		0.068	109.15
0.029	16.88		0.069	113.53
0.030	20.69		0.070	107.72
0.031	14.24		0.071	112.04
0.032	35.34		0.072	120.60
0.033	32.21		0.073	127.08
0.034	22.52		0.074	128.41
0.035	22.72		0.075	128.46
0.036	34.14		0.076	123.03
0.037	39.65		0.077	131.32
0.038	49.97		0.078	129.26
0.039	43.65		0.079	140.73
0.040	33.63		0.080	126.05
0.041	43.67		0.081	140.28
0.042	46.12		0.082	138.63
0.043	47.85		0.083	143.68
0.044	58.72		0.084	134.71
0.045	53.93		0.085	140.83
0.046	54.84		0.086	140.35
0.047	68.18		0.087	150.50
0.048	54.17		0.088	156.06
0.049	63.19		0.089	156.62
0.050	67.53		0.090	178.98
0.051	65.37		0.091	160.56
0.052	89.13		0.092	154.60
0.053	82.43		0.093	171.39
0.054	91.49		0.094	171.64
0.055	106.44		0.095	173.81
0.056	86.49		0.096	180.18
0.057	75.37		0.097	174.17
0.058	94.19		0.098	181.31
0.059	98.72		0.099	175.71
0.060	95.22		0.100	192.20

Estimating POD Limit, $y_{90/95\ \text{POD}}$

The $a_{90/95}$ value for the dataset is estimated as 0.047 inch using MIL-HDBK-1823A methods given a threshold of $y_{dec} = 40$. LS-POD is expected to be conservative, therefore a larger defect size of 0.060 inch was chosen for the simulated LS-POD experiment. The following process can be applied to any defect size.

Using the estimated model parameters, 10 inspections at 0.060 inch were simulated in Table B.5 by calculating the mean response and adding a random residual from a normal distribution with a mean of zero and standard deviation of 7.30.

Table B.5. Simulated Flaw Signals for Raised Head Fastener Inspection

	defect size 0.060”		
defect number	mean response from model	random response	total response
1	91.68	-12.70	78.98
2	91.68	8.86	100.54
3	91.68	1.02	92.70
4	91.68	-7.29	84.39
5	91.68	1.61	93.29
6	91.68	-4.81	86.87
7	91.68	-3.71	87.97
8	91.68	10.18	101.86
9	91.68	17.36	109.04
10	91.68	-2.01	89.67

The sample mean and standard deviation of the 10 flaw responses are estimated according to Appendix A as $\bar{y}_F = 92.4$ and $s_F = 9.04$. From Appendix C, for sample size of 10, $k_{1_F} = 2.355$, and $y_{90/95\ \text{POD}}$ is estimated as

$$y_{90/95\ \text{POD}} = 92.4 - 2.355 * 9.04 = 71.1$$

Therefore, this represents the maximum decision threshold that maintains 90% POD with 95% confidence. Signals above $y_{90/95\ \text{POD}} = 71.1$ could called as a hit by the NDE inspection.

Estimating POF Limit, $y_{1/95\ \text{POF}}$

The noise data were not explicitly measured in the original experiment. Based on NDE engineering judgment, the noise data for unflawed specimen inspections was modeled as lognormal, with mean $m = 5$ and standard deviation $v = 5$. To simulate the noise measurements, the parameters of the lognormal distribution were converted to the mean, μ, and standard deviation, σ, of a normal distribution as follows.

$$\mu = \log\left(\frac{m^2}{\sqrt{v + m^2}}\right), \sigma = \sqrt{\log\left(1 + \frac{v}{m^2}\right)}$$

$$\mu = \log\left(\frac{5^2}{\sqrt{5+5^2}}\right), \sigma = \sqrt{\log\left(1+\frac{5}{5^2}\right)}$$

$$\mu = 1.52, \sigma = 0.43$$

Therefore, the log of the noise data is simulated from a normal distribution with $\mu = 1.52$ and $\sigma = 0.43$. Forty simulated noise measurements are shown in Table B.6 in the form of as-measured signals and log transformed signals.

Table B.6. Simulated Noise Signals

noise signal	log(noise signal)		noise signal	log(noise signal)
4.186	1.432		7.139	1.966
3.625	1.288		3.912	1.364
5.649	1.731		3.343	1.207
3.535	1.263		10.905	2.389
3.154	1.149		4.523	1.509
9.975	2.300		5.300	1.668
9.259	2.226		7.075	1.957
8.230	2.108		5.645	1.731
3.529	1.261		12.332	2.512
8.083	2.090		3.858	1.350
10.868	2.386		6.466	1.867
2.122	0.752		4.882	1.585
10.767	2.376		4.362	1.473
4.989	1.607		2.478	0.907
5.494	1.704		2.268	0.819
5.380	1.683		8.175	2.101
5.531	1.710		5.325	1.672
2.899	1.064		11.376	2.431
3.456	1.240		4.375	1.476
4.097	1.410		6.366	1.851

A histogram of the 40 simulated unflawed measurements is shown in Figure B.6.

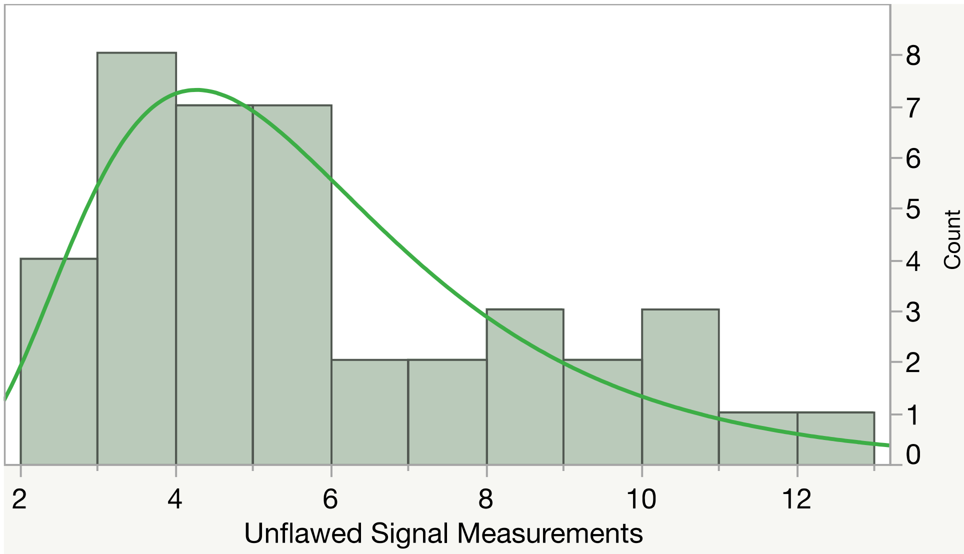

Figure B.6. Distribution of unflawed measurements for raised head fastener inspections with an overlayed lognormal density.

The noise limit is found by analyzing the logarithm of the noise measurements. The mean and standard deviation of these 40 measurements are 1.67 and 0.459. From Appendix D, for a sample size of 40, $k_{1_N} = 2.941$, and the transformed $\log\left(y_{1/95\text{ POF}}\right)$ is estimated as

$$\log\left(y_{1/95\text{ POF}}\right) = 1.67 + 2.941 * 0.459 = 3.015$$

And now $y_{1/95\text{ POF}} = \exp(3.015) = 20.39$

Therefore, signals below $y_{1/95\text{ POF}} = 20.39$ should not be called as a hit by the NDE inspection to maintain a POF of 1% with 95% confidence.

Interpretation of the LS-POD Results

Figure B.7 illustrates the noise and flaw distribution and the $y_{90/95\text{ POD}}$ and $y_{1/95\text{ POF}}$ limits for this example. Based on the guidance provided in Section 2.5.1, the result that $y_{90/95\text{ POD}} \geq y_{1/95\text{ POF}}$ satisfies the first criteria of a successful demonstration at 90/95 POD and 1/95 POF. A decision threshold value between the POF and POD limits, where $20.39 \leq y_{dec} \leq 71.1$ is considered acceptable for the inspection of the flight hardware.

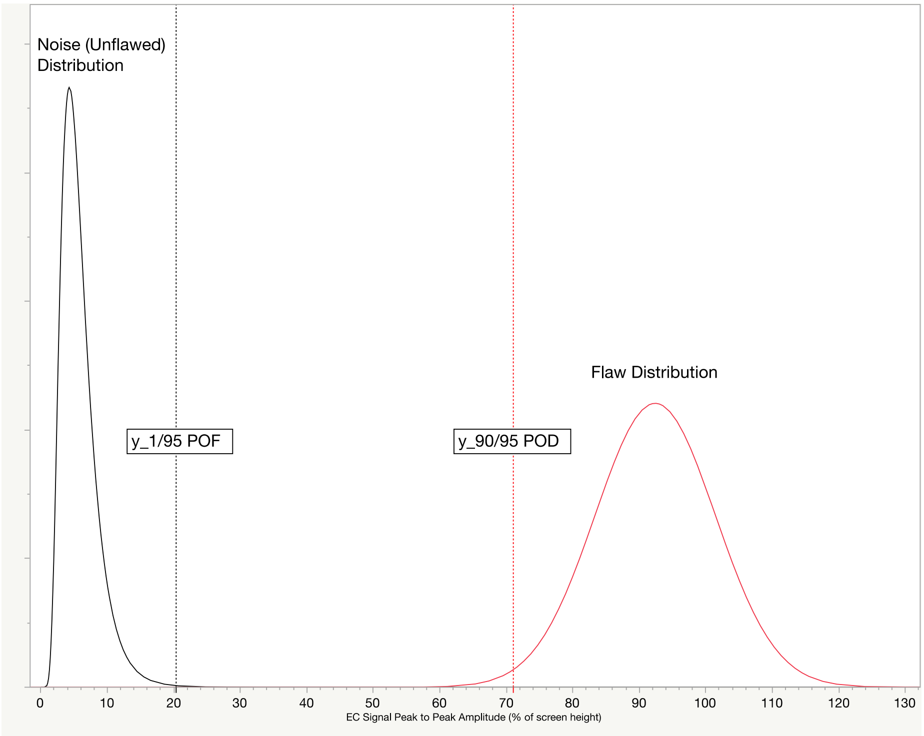

Figure B.7. Noise (unflawed) and flaw signal distribution for raised head fastener demonstration.

References

David Forsyth, Mark Gehlen, Jeff Guthrie, Mark Keiser, Ronald Kent, Michael Morgan, Carlos Pairazaman, Darren Stamper, Damaso Carreon, "The Air Force Nondestructive Inspection Improvement Program," United States Air Force Aircraft Structural Integrity Program Conference, San Antonio, 30 Nov – 02 Dec 2010. http://meetingdata.utcdayton.com/agenda/asip/2010/proceedings/presentations/P4213.pdf

Appendix C. Table of k_1 values for flaw measurements to provide 90 percent probability/95-percent confidence for calculating POD for samples sizes ranging from 10 to 30 based on formulas provided in Appendix A.

Sample Size	k_{1_F}(90/95)
10	2.355
11	2.275
12	2.210
13	2.155
14	2.109
15	2.068
16	2.033
17	2.002
18	1.974
19	1.949
20	1.926
21	1.905
22	1.886
23	1.869
24	1.853
25	1.838
26	1.824
27	1.811
28	1.799
29	1.788
30	1.777

Appendix D. Table of k_1 values for unflawed measurements that provide 1 percent probability/95-percent confidence for calculating POF with sample sizes ranging from 40 to 60 based on formulas provided in Appendix A.

Sample Size	k_{1_N} (1/95)
40	2.941
41	2.932
42	2.923
43	2.914
44	2.906
45	2.898
46	2.890
47	2.883
48	2.876
49	2.869
50	2.862
51	2.856
52	2.850
53	2.844
54	2.838
55	2.833
56	2.827
57	2.822
58	2.817
59	2.812
60	2.807

Appendix E. Analytical Comparison of Limited Sample POD to MIL-HDBK-1823A

This appendix provides an analytical comparison of Limited Sample Probability of Detection (LS-POD) signal limit, $y_{90/95\ \mathrm{POD}}$, to the $a_{90/95}$ flaw size resulting from a MIL-HDBK-1823A POD study, and shows that LS-POD is expected to be conservative relative to MIL-HDBK-1823A's. Throughout this algebraic derivation, the estimated parameters can be considered as their expected values, even though the notation is not explicit, to illustrate the relationship rather than focus on the statistical aspects.

In MIL-HDBK-1823A, a linear model of $\hat{a}\ vs\ a$ provides an estimated intercept of $\hat{\beta}_0$, slope of $\hat{\beta}_1$, and regression modeling error of $\hat{\tau}$. It can be shown that the $a_{90/95}$ flaw size is estimated by

$$a_{90/95} = \frac{1}{\hat{\beta}_1}\left(y_{dec} - \hat{\beta}_0 + k_{1_{MH}} * \ \hat{\tau}\right)$$

where y_{dec} is the specified decision threshold, and $k_{1_{MH}}$ is the tolerance interval multiplier associated with $(m_{MH} - 2)$ degrees of freedom for m_{MH} measurements across multiple flaw sizes. The subscript MH signifies estimates from a MIL-HDBK-1823A POD study, and a subscript of LS will indicate results from a LS-POD demonstration test.

Without loss of generality, assume that $\hat{\beta}_0$= 0 and $\hat{\beta}_1$= 1, therefore

$$a_{90/95} = y_{dec} + k_{1_{MH}} * \ \hat{\tau}$$

Rearranging to solve for the decision threshold, y_{dec}, results in

$$y_{dec} = a_{90/95} - k_{1_{MH}} * \ \hat{\tau} \tag{1}$$

In the LS-POD method described in Appendix A, the 90/95 POD signal limit is found by

$$y_{90/95\ \mathrm{POD}} = \ \bar{y}_F - k_{1_{LS}} * s_F \tag{2}$$

where $\bar{y}_F$ and s_F are the sample mean and standard deviation of m_{LS} flaw signal measurements, and $k_{1_{LS}}$ is a tolerance interval factor based on m_{LS} flaw signal measurements of the target flaw size.

If the $y_{90/95\ \mathrm{POD}}$ POD limit from a LS-POD demonstration experiment were equal to the decision threshold, y_{dec}, from MIL-HDBK-1823A then, $y_{90/95\ \mathrm{POD}} = y_{dec}$.

Therefore, combining Equations 1 and 2, results in the following expression

$$\bar{y}_F - k_{1_{LS}} * s_F = a_{90/95} - k_{1_{MH}} * \ \hat{\tau} \tag{3}$$

where the left side of this expression is in units of the signal and the right side is in units of the flaw size. However, since $\hat{\beta}_0$= 0 and $\hat{\beta}_1$= 1 the units are removed, and they can be numerically equal.

LS-POD demonstration experiments are designed to require a smaller sample size than a MIL-HDBK-1823A POD study, therefore $n_{LS} < n_{MH}$ is expected. Since the k_1 multipliers are inversely proportional to sample size, it follows that

$$k_{1_{LS}} > k_{1_{MH}}$$

Therefore, if $s_F = \hat{\tau}$ (see footnote 1), and allowing $s_F = 1$, without loss of generality, Equation 3 becomes

$$\bar{y}_F - k_{1_{LS}} = a_{90/95} - k_{1_{MH}}$$

and rearranging, the expression becomes

$$\bar{y}_F - a_{90/95} = k_{1_{LS}} - k_{1_{MH}}$$

where $k_{1_{LS}} - k_{1_{MH}} > 0$, therefore by inspection

$$\bar{y}_F > a_{90/95}$$

While $\bar{y}_F$ is the average signal of the flaw distribution, by letting $\hat{\beta}_1 = 1$ it is also equal to the average flaw size of LS-POD specimens. This result shows that the target flaw size for a LS-POD demonstration is required to be larger than the $a_{90/95}$ flaw size from a MIL-HDBK-1823A POD study given the same decision threshold, $y_{90/95\ \mathrm{POD}} = y_{dec}$.

The signal decision threshold tends to be limited by the unflawed, noise distribution, and therefore it is assumed that it cannot be lowered to detect the $a_{90/95}$ flaw size in a LS-POD demonstration.

This derivation emphasizes the small sample penalty of LS-POD, and it shows that LS-POD is expected to be conservative relative to a MIL-HDBK-1823A POD study.

Footnote:

[1] $\hat{\tau}$ is estimated from the regression modeling of $\hat{a}\ vs\ a$ in a MIL-HDBK-1823A POD study across multiple flaw sizes, while s_F is estimated from a LS-POD demonstration from a single target flaw size. Technically, they represent different types of variability; however, for purposes of this derivation, it is assumed that extracting a subset of signal measurements at a single flaw size from a MIL-HDBK-1823A POD study can be treated as a LS-POD demonstration experiment and that the residual variance is representative of the signal variability at that target flaw size. This is demonstrated in the Example 1 dataset from MIL-HDBK-1823A in Appendix B, and it numerically confirms the results of this derivation, namely that $\bar{y}_F > a_{90/95}$.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
02/02/2022	Technical Memorandum	

4. TITLE AND SUBTITLE Guidebook for Limited Sample Probability of Detection (LS-POD) Demonstration for Signal-Response Nondestructive Evaluation (NDE) Methods	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Koshti, Ajay; Parker, Peter A.; Forsyth, David S.; Suits, Michael W.; Walker, James L.; Prosser, William H.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER 869021.01.23.01.01
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Langley Research Center Hampton, VA 23681-2199	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001	**10. SPONSOR/MONITOR'S ACRONYM(S)** NASA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) NASA/TM-20210018515/Corrected

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category Mechanical Engineering
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This document provides guidance for performing limited sample probability of detection (LS-POD) demonstration testing with smaller numbers of flawed specimens than the 29 required for the Point Estimate POD method referenced in NASA Standard 5009B. This document is specifically applicable to NASA programs and projects where NASA Special NDE methods are used to inspect fracture critical human spaceflight metallic hardware when the minimum number of flawed specimens required for the Point Estimate POD method referenced in NASA Standard 5009B (2019) are not available.

15. SUBJECT TERMS
Limited Sample Probability of Detection; Nondestructive Evaluation; Demonstration Testing; Probability of False Calls; Guidebook

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	44	(443) 757-5802