AFRL-ML-WP-TR-2001-4010

PROBABILITY OF DETECTION (POD) ANALYSIS FOR THE ADVANCED RETIREMENT FOR CAUSE (RFC)/ENGINE STRUCTURAL INTEGRITY PROGRAM (ENSIP) NONDESTRUCTIVE EVALUATION (NDE) SYSTEM DEVELOPMENT
VOLUME 1 – POD ANALYSIS

ALAN P. BERENS

UNIVERSITY OF DAYTON
RESEARCH INSTITUTE
300 COLLEGE PARK
DAYTON, OH 45469-0120

JANUARY 2000

FINAL REPORT FOR PERIOD 29 SEPTEMBER 1995 – 31 DECEMBER 1999

Approved for public release; distribution unlimited.

MATERIALS AND MANUFACTURING DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7750

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
00012000	N/A	-

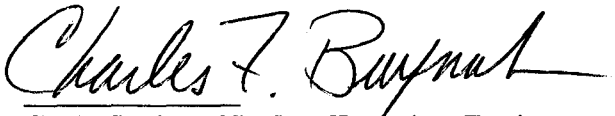
Title and Subtitle Probability of Detection (POD) Analysis for the Advanced Retirement for Cause (RFC)/Engine Structural Integrity Program (ENSIP) Nondestructive Evaluation (NDE) System-Volume 1: POD Analysis	**Contract Number**
	Grant Number
	Program Element Number
Author(s) Berens, Alan P.	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) University of Dayton Research Institute 300 College Park Dayton, OH 45469-0120	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) Materials and Manufacturing Directorate Air Force Research Laboratory Air Force Materiel Command Wright-Patterson AFB, OH 45433-7750	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes The original document contains color images.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 99	

NOTICE

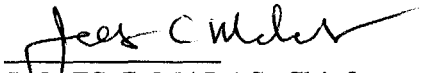
WHEN GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA ARE USED FOR ANY PURPOSE OTHER THAN IN CONNECTION WITH A DEFINITELY GOVERNMENT-RELATED PROCUREMENT, THE UNITED STATES GOVERNMENT INCURS NO RESPONSIBILITY OR ANY OBLIGATION WHATSOEVER. THE FACT THAT THE GOVERNMENT MAY HAVE FORMULATED OR IN ANY WAY SUPPLIED THE SAID DRAWINGS, SPECIFICATIONS, OR OTHER DATA, IS NOT TO BE REGARDED BY IMPLICATION OR OTHERWISE IN ANY MANNER CONSTRUED, AS LICENSING THE HOLDER OR ANY OTHER PERSON OR CORPORATION, OR AS CONVEYING ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY IN ANY WAY BE RELATED THERETO.

THIS REPORT IS RELEASABLE TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS). AT NTIS, IT WILL BE AVAILABLE TO THE GENERAL PUBLIC, INCLUDING FOREIGN NATIONS.

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.



CHARLES F. BUYNAK, Project Engineer
Nondestructive Evaluations Branch
Metals, Ceramics & NDE Division



JAMES C. MALAS, Chief
Nondestructive Evaluations Branch
Metals, Ceramics & NDE Division

GERALD J. PETRAK, Assistant Chief
Metals, Ceramics & NDE Division
Materials & Manufacturing Directorate

IF YOUR ADDRESS HAS CHANGED, IF YOU WISH TO BE REMOVED FROM OUR MAILING LIST, OR IF THE ADDRESSEE IS NO LONGER EMPLOYED BY YOUR ORGANIZATION, PLEASE NOTIFY, AFRL/MLLP, WRIGHT-PATTERSON AFB OH 45433-7817 AT (937) 255-9819 TO HELP US MAINTAIN A CURRENT MAILING LIST.

COPIES OF THIS REPORT SHOULD NOT BE RETURNED UNLESS RETURN IS REQUIRED BY SECURITY CONSIDERATIONS, CONTRACTUAL OBLIGATIONS, OR NOTICE ON A SPECIFIC DOCUMENT.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	JANUARY 2000	Final, 09/29/1995 – 12/31/1999

4. TITLE AND SUBTITLE
PROBABILITY OF DETECTION (POD) ANALYSIS FOR THE ADVANCED RETIREMENT FOR CAUSE (RFC)/ENGINE STRUCTURAL INTEGRITY PROGRAM (ENSIP) NONDESTRUCTIVE EVALUATION (NDE) SYSTEM DEVELOPMENT

VOLUME 1 – POD ANALYSIS

5. FUNDING NUMBERS
C: F33615-95-C-5242
PE: 63112F
PN: 3153
TN: 00
WU: 19

6. AUTHOR(S)
ALAN P. BERENS

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF DAYTON
RESEARCH INSTITUTE
300 COLLEGE PARK
DAYTON, OH 45469-0120

8. PERFORMING ORGANIZATION REPORT NUMBER
UDR-TR-2000-00007

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
MATERIALS AND MANUFACTURING DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7750
POC: Charles Buynak, AFRL/MLLP, (937) 255-9807

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-ML-WP-TR-2001-4010

11. SUPPLEMENTARY NOTES
This is Volume 1 of 3 Volumes.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

In 1995, the US Air Force initiated a multifaceted program to upgrade the Eddy Current Inspection System (ECIS) system. Planned changes to the ECIS system included updating a) the eddy current instrument, b) the station computer and its operating system, and c) the robotics controller. These system changes were to be "drop-in" to avoid the costs of repeating a complete capability demonstration program for ongoing engine inspections. In addition, a new approach to calibration was demonstrated and a new Probability of Detection (POD) computer program was written for the analysis of data from capability demonstrations. This report presents the results of the study whose objectives were updating the POD analysis program and evaluating the data collected to demonstrate the drop-in compatibility of the upgraded and original ECIS systems. An additional task was later added to analyze data from an experiment designed to validate two proposed methods for inferring the POD of a geometry/material combination using inspection results from other geometry/material combinations. While the primary focus of this study was POD analysis for the RFC/ENSIP application with its highly automated ECIS, the POD analysis methods are applicable to a broad range of nondestructive evaluation (NDE) systems.

14. SUBJECT TERMS	15. NUMBER OF PAGES
retirement for cause, engine structural integrity, probability of detection, eddy current	102
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

NSN 7540-01-280-5500 — **Standard Form 298 (Rev. 2-89)** Prescribed by ANSI Std. Z39-18 298-102

Table of Contents

Section **Page**

List of Figures

List of Tables

Acronyms, Abbreviations, and Symbols

AFRL	Air Force Research Laboratory
AHAT	Computer program for POD analysis from quantified signal responses
ECIS	Eddy Current Inspection System
EDM	Electrical Discharge Machined
ENSIP	Engine Structural Integrity Program
NDE	Nondestructive Evaluation
P/F	Computer program for POD analysis from hit/miss inspection responses
POD	Probability of Detection
POD(a)	Probability of Detection of flaws of size a
RFC	Retirement For Cause
a	Crack size
$\hat{a}$ or ahat	Response of NDE system to flaw
A*	Anderson - Darling test statistic for normality
a_{90}	"90 percent detectable crack size, POD($a90$) = 0.9"
$a_{90/50}$	Best estimate of 90 percent detectable crack size (about 50 percent confidence)
$a_{90/95}$	Estimate of 90 percent detectable crack size with 95 percent confidence
$\hat{a}_{dec}$	Response signal threshold for crack detection
ahat or $\hat{a}$	Response of NDE system to flaw
a_i	Size of crack i
$\hat{a}_i$	Response of NDE system to flaw i
a_{NDE}	Reliably detected crack size
a_p	Crack size for which POD(ap) = 100 p
$a_{p/q}$	Estimate of p percent detectable crack size with q percent confidence
B_0, B_1	Slope and intercept of linear relation between log â and log a
C	Differential random effect on â due to calibration blocks
χ^2	Equal variance test statistic - χ^2 distribution
c_i	Differential random effect due to crack i
CP	Differential random effect on $\hat{a}$ due to the interaction of calibration blocks and probes
$\boldsymbol{d}$	Difference between predicted and actual â or partial derivative operator
$\boldsymbol{d_i}$*	Sum of random effe cts due to c_i and $p_j(i)$
$\boldsymbol{D}X$	Error in estimate of parameter X
F	Lack of fit test statistic - F distribution
$\boldsymbol{F}(\ )$	Standard cumulative normal distribution, i.e., zero mean and unit standard deviation
$f(a)$	Functional relation between a and $\hat{a}$
$g(\hat{a})$	Monotonic transformation of $\hat{a}$
$G_i M_j$	Specimen comprised of Geomtry i and Material j
$h(a)$	Monotonic transformation of a
i	Index parameter
j	Index parameter
k	Number of inspections of a flaw
$L(\boldsymbol{q})$	Likelihood of $\boldsymbol{q}$ as a function of observed inspection results

ln	Base e logarithm
m	Location parameter of POD(a) model, exp(μ) is 50% detectable crack size
$\hat{\mu}$	Maximum likelihood estimate of m
$M(x_p)$	Mean of x_p
n	Number of flaws
$\boldsymbol{P}$	Product
P	Differential random effect on â due to probes
$p_j(i)$	Differential random effect due to probe j and recalibration on inspection of crack i
$\boldsymbol{q}$	Vector of parameters of a general POD(a) function
$\boldsymbol{s}$	Steepness parameter of POD(a) model
$\hat{\boldsymbol{s}}$	Maximum likelihood estimate of $\boldsymbol{s}$
$\boldsymbol{s}_{\boldsymbol{d}}$	Standard deviation of $\boldsymbol{d}$
SD(Xp)	Standard deviation of Xp
$\boldsymbol{s}_p$	Pooled standard deviation of residuals from multiple $\hat{a}$ recording of single crack
$\boldsymbol{s}_x^2$	Variance of arbitrary parameter x
V_{ii}	Variance of the maximum likelihood estimates of parameter i
V_{ij}	Covariance of the maximum likelihood estimates of parameters i and j
X	Transformed crack size, e.g., $X = h(a)$
X_p	pth percentile of normal distribution of X
Y	Transformed inspection response, e.g., $Y = g(\hat{a})$
Z_i	Hit or miss response for inspection i, $Z_i = 0$ implies miss (fail or no find), $Z_i = 1$ implies hit (pass or find)
z_p	pth percentile of a standard normal distribution

Foreword

This technical report was prepared by the University of Dayton Research Institute for the Materials and Manufacturing Directorate of the Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio. The work was performed under Contract Number F33615-95-C-5242, with Mr. Charles F. Buynak (AFRL/MLLP) as the Air Force project engineer. The technical effort was performed between 29 September 1995 and 31 December 1999, with Dr. Alan P. Berens of the University of Dayton Research Institute as the principal investigator.

The final report of this work comprises three volumes. Volume 1 presents a description of changes made to the probability of detection (POD) analysis program of Mil-HDBK-1823 and the statistical evaluation of modifications that were made to version 3 of the Eddy Current Inspection System (ECIS v3). Volume 2 contains the *Users Manual* for the version 3 update of the POD program. The results of a separate study for predicting POD from specimens of like geometry and materials are presented in Volume 3.

Section 1
Introduction

The Retirement For Cause/Engine Structural Integrity Program (RFC/ENSIP) approach to engine life management and safety is centered on the projected growth of the largest flaw that might be in a structure at the beginning of a period of operational usage [1,2]. Because the intervals between required maintenance actions are inversely proportional to the size of this largest potential flaw, there is a strong economic incentive to reliably detect ever smaller flaws. During the 1980's, the U.S. Air Force developed and implemented a highly automated, eddy current inspection system (ECIS) for detecting cracks in engine components [3]. Since the detection of the small cracks of interest is stochastic in nature, an approach for quantifying inspection capability in terms of crack size was also developed and promulgated during this period [4,5].

In 1995, the U.S. Air Force initiated a multifaceted program to upgrade the ECIS. Planned changes to the ECIS include updating a) the eddy current instrument, b) the station computer and its operating system, and c) the robotics controller. These system changes were to be "drop-in" to avoid the costs of repeating a complete capability demonstration program for ongoing engine inspections. In addition, a new approach to calibration was demonstrated, and a new POD computer program was written for the analysis of data from capability demonstrations. This report presents the results of the study whose objectives were updating the POD analysis program and evaluating the data collected to demonstrate the drop-in compatibility of the upgraded and original ECISs. An additional task was later added to analyze data from an experiment designed to validate two proposed methods for inferring the POD of a geometry/material combination using inspection results from other geometry/material combinations.

While the primary focus of this study was POD analysis for the RFC/ENSIP application with its highly automated ECIS, the POD analysis methods are applicable to a broad range of nondestructive evaluation (NDE) systems. Accordingly, the report discusses POD analyses in this broader context.

1.1 Overview of POD Demonstrations

The damage tolerance philosophy for ensuring structural integrity focuses on predictions of the growth of cracks at critical locations in structural details [2,6]. In particular, deterministic safety analyses are performed which demonstrate that the most severe crack that might be in a fracture-critical component at the beginning of a usage interval will not grow to critical size before being detected and repaired. After an in-service maintenance action (inspection and repair when necessary), the assumed severe crack size, say a_{NDE}, is defined to be the largest crack that might be missed at the inspection. Smaller a_{NDE} values result in the benefits of longer intervals between inspections, but cracks that are smaller than a specific threshold should not be detected because detection and repair of such nonthreatening cracks is not economically sensible. Thus, the objective in the design of an NDE system is to reliably detect all cracks greater than a_{NDE} but not to obtain crack indications at locations with no cracks or no significant cracks.

Although the response to an inspection stimulus is dependent on crack size, the magnitude of the response is not determined by size alone. Many other factors are also correlated with the response signal. Not all of these factors can be controlled or accounted for by the inspection system and, in fact, some are inherently random in routine applications of an inspection system. This uncertainty in the detection of cracks leads to characterizing the capability of an NDE system in terms of the probability of detection as a function of crack size, POD(a).

Four methods are currently being used or considered in the estimation of a_{NDE} or POD(a) functions for inspection systems: a) engineering judgement, b) theoretical modeling, c) past inspection results, and d) demonstration experiments. This study considers only the demonstration experiment approach in which specimens with known crack sizes are inspected. The specimens are assumed to be representative of the real inspections, and the parameters of the POD(a) model are estimated from the inspection results. The reliably detected crack size, a_{NDE}, for an NDE system is defined in terms of a crack size for which there is a high probability of detection. Usually POD(a_{NDE}) = 0.90, and this crack size is often designated as the a_{90} value. Since there is statistical uncertainty in the parameter estimates, there is also statistical uncertainty in the estimate of a_{NDE}. To account for this uncertainty, a statistically based upper confidence limit can be placed on the a_{NDE} estimate. Usually, a 95 percent confidence limit is used for this characterization of inspection capability, and a_{NDE} is defined as the crack size for which there is 95 percent confidence that POD(a_{NDE}) ≥ 0.9. This crack size characterization has been designated as the $a_{90/95}$ crack size for an inspection and is also referred to as the 90/95 crack size. Similarly, it has become customary to refer to the best estimate of the 90 percent detectable crack as $a_{90/50}$ or the 90/50 crack size. Note that the $a_{90/50}$ and $a_{90/95}$ values are not characteristic properties of an NDE system, but rather are calculated from the particular random results of the capability experiment. If the capability experiment were repeated, different $a_{90/50}$ and $a_{90/95}$ values would be obtained. Note also that, although a_{NDE} has been commonly characterized by $a_{90/50}$ or $a_{90/95}$, other values of POD and confidence levels could be used. This report was written with an emphasis on estimating $a_{90/50}$ or $a_{90/95}$, but the updated computer program permits any choice of POD and confidence levels of 90, 95, and 99 percent.

Inspection results are recorded in two different formats, and the format determines the analysis method to be used in modeling the POD(a) function. When the results of an inspection are expressed only in terms of whether or not a crack was detected, the data are known as find/no find, pass/fail, or hit/miss. Such dichotomous inspection results are represented by the data pair (a_i, Z_i), where a_i is the size of the i^{th} crack and Z_i represents the outcome of the inspection of the i^{th} crack; Z_i = 1 for the crack being hit (find or pass) and Z_i = 0 for the crack being missed (no find or fail). Examples of such data would be the results of visual, magnetic particle, or fluorescent penetrant inspection, or any inspection for which the magnitude of the response to the inspection stimulant was not recorded. The POD(a) analysis for data of this nature is often called pass/fail or hit/miss analysis. Maximum likelihood estimates of the parameters of the POD(a) model are obtained from the (a_i, Z_i) data. Asymptotic properties of the maximum likelihood estimates are used to calculate the confidence bound on the estimate of a_{NDE}.

When the results of the inspection are based on the quantified magnitude of a response to the NDE stimulus and the response is recorded, the POD(a) function can be estimated from the

statistical scatter in the response magnitudes as a function of crack size. The data pair comprising size and signal response are designated as (a_i, $\hat{a}_i$), in which $\hat{a}_i$ is the response to the NDE stimulus for the i^{th} crack. If $\hat{a}_i$ is greater than a preset threshold, $\hat{a}_{th}$, a crack is indicated. Data of this nature are often referred to as $\hat{a}$ versus a (*ahat* versus *a*).The data from the automated ECIS are of this nature, and data from ultrasonic and fluorescent penetrant inspections have also been recorded and analyzed in the $\hat{a}$ versus a format. The parameters of the POD(a) function are estimated from the scatter in $\hat{a}$ values about the mean response to cracks of size a. Again, maximum likelihood estimates are used to estimate the parameters and to place confidence bounds on the estimate of a_{NDE} when desired.

The application of maximum likelihood to the estimation of the parameters of the POD(a) model and a_{NDE} are presented in detail in Mil-HDBK-1823 [4], Berens [5], and Petrin, et al. [7]. These details will not be repeated in this report. Only the analysis changes and additions will be discussed.

1.2 Update POD Program

The assessment of NDE capability through a demonstration experiment requires careful technique in the planning and execution of the inspections, as well as in the statistical analysis of the test results. The current test protocol and analysis methodology for assessing NDE system capability in aircraft engines evolved during the 1980's and is summarized in Mil-HDBK-1823 [4]. The same material can be found in Petrin, et al. [7]. Rigid adherence to the dictates of references [4 and 7] will lead to a correct NDE capability characterization in terms of POD(a). However, the computer code for performing the POD analyses is based on the outdated computer technology of the early 1980's. Current computer capabilities permit more automated analyses and the direct generation of report quality output.

A prime objective of the study reported herein was to generate an updated POD analysis program. To set the analysis framework, section 2 of this report addresses the demonstration experiment approach to evaluating the POD capability of any NDE system and discusses the design of demonstration experiments and the analysis of the resulting data. Section 3 presents the modifications that were incorporated in the new POD analysis computer program. The *Users Manual* for the new version of the computer is presented as Volume 2 of this report.

1.3 Evaluate ECIS Modifications

Significant changes are planned for the ECIS RFC/NDE system as a result of improvements to components that were developed in the last four years. In particular, changes to the eddy current instrument, the station computer/operating system, and the robotics controller are in the process of being incorporated in the ECIS. One of the criteria required for potential planned changes to the system was that the changes could be incorporated using the pre-existing thresholds of the original system. This criterion was interpreted to mean that the a_{90} values for the modified system would be equivalent to those of the original system without any changes in $\hat{a}$ detection thresholds. The same criterion was imposed on the newly developed method for calibrating the system.

The evaluation of the data collected to verify the drop-in compatibility of the ECIS system modifications is presented in section 4. The compatibility of each of the modifications was individually tested by a comparison of the $\hat{a}$ values from individual cracks and a comparison of a_{90} values as a function of detection thresholds. Similar analyses were performed to demonstrate the compatibility of new approach to calibration and this evaluation is also included in section 4.

1.4 Inferring POD(*a*) Through Material Correlation

The specimens used in POD demonstrations are assumed to be representative of the real components that will be inspected by the NDE system. These specimens are generally not real parts but rather are intended to mimic the parts in the important inspection response parameters. Specimen geometry and material are known to influence the NDE signal response and most inspection capability demonstrations have been performed on specimens of equivalent geometry and material combinations. The cost of producing sets of specimens with fatigue cracks covering a range of sizes is high. Significant savings are possible given a valid method for obtaining a_{NDE} values without having to first manufacturer sets of specimens for all possible combinations of material and geometry.

A scenario is envisioned in which POD(*a*) for a particular geometry - material combination, say G_1M_1, can be obtained from the inspection of specimens of a like geometry but different material, say G_1M_2, and a different geometry but the same materials, say G_2M_1 and G_2M_2. The practical problems in conducting such evaluations and two proposed methods for inferring the POD(*a*) function for a missing combination given three of four sets of data were evaluated as part of this program. Since this small study was not a prime objective of the program, the evaluation results are presented in Volume 3.

Section 2
Capability Demonstrations

The POD capability of an NDE system is typically estimated through a capability demonstration program. The concept is to mimic the real inspection as closely as possible on representative specimens that contain cracks spanning the range of increase of the POD(a) function. A comprehensive description for the execution of such a demonstration program and the analysis of the resulting data is presented in Mil-HDBK-1823 [4]. The analysis of data from an NDE demonstration is based on maximum likelihood estimates of the parameters of the POD(a) model and the asymptotic properties of such estimates. The mathematical details of these analyses are fully presented in Mil-HDBK-1823 [4] and Berens [5]. This section briefly reviews the design and execution of a generic capability demonstration and presents changes to the analyses based on insights that have been gained since Mil-HDBK-1823 was written.

2.1 Demonstration Plan

An NDE reliability demonstration comprises the execution of a test matrix of inspections on a set of specimens with known flaw locations and sizes. The inspection results, either $\hat{a}$ or hit/miss, are then analyzed to estimate the POD(a) function and a_{NDE} for the inspection application. The specimens are inspected under a test protocol that simulates as closely as practical the actual application conditions. Establishing test protocols for eddy current, fluorescent penetrant, ultrasonic and magnetic particle inspection systems are discussed in Mil-HDBK-1823 [4]. This report addresses only the analysis of the resulting data that is governed by the nature of the inspection result ($\hat{a}$ or hit/miss) and the experimental design of the demonstration.

The objectives and costs of an NDE demonstration determine the matrix of inspections to be performed. From the analysis viewpoint, there are two major categories of concerns that must be addressed in establishing the experimental design. These are as follows: a) the generality of inferences that can be made from the controlled and uncontrolled inspection and material parameters, and b) the number and sizes of flaws and the number of unflawed inspection sites in the specimens. These topics are addressed in the following subsections.

2.1.1 Controlled and Uncontrolled Factors

The demonstration of NDE capability is both a consumer or and quality concern. The primary objective of such demonstrations for a particular application is to estimate the POD(a) function and, consequently, a_{NDE}. For damage tolerance considerations, a_{NDE} is commonly accepted to be the crack sizes designated as a_{90} or $a_{90/95}$. The a_{90} crack size is defined as the size for which POD(a_{90}) = 0.90 and $a_{90/95}$ is the upper (conservative) 95 percent confidence bound on the estimate of a_{90}. NDE reliability experiments have also been conducted to optimize the inspection protocol and to ensure process control. System optimization with respect to POD(a) would have the objective of determining system

configurations that produce acceptable a_{90} or $a_{90/95}$ values. The design of system optimization programs is of a different character and beyond the current scope.

To demonstrate capability for an application, it is assumed that: a) the complete protocol for conducting the inspection is well defined for the application, b) the inspection process is under control, and c) all other factors which introduce variability in an inspection decision are reasonably representative of the application. The representativeness of these other factors limits the scope of the POD(a) characterization and is addressed by controlling the factors during the inspection or by randomly sampling the factors to be used in the demonstration. The methods of accounting for these factors are important aspects of the statistical design of the demonstration and significantly influence the statistical properties of the estimates of the POD(a) function parameters.

The important classes of the factors that introduce variation in crack detectability are as follows:

a) the inherent degree of repeatability of the magnitude of the NDE signal response when a specific crack is independently inspected many times with all controllable factors held constant
b) the material and geometrical properties of the specimens and the differences in the physical properties of flaws of nominally identical "size"
c) the variation introduced by different hardware components in the inspection system
d) the summation of all the human factors associated with the particular population of inspectors that might be used in the application.

The effects of these factors are present in every NDE reliability demonstration, and they should be explicitly considered in the design of the demonstration and the interpretation of the results.

2.1.1.1 Inherent Variability

Little can be done about the variation of the response to the NDE excitation at the demonstration stage when inspections are repeated under fixed conditions. This variation might be reduced if the system was modified or better optimized, but that is a different objective. Repeat inspections under identical conditions will provide a measure of the inherent variability that is a lower bound on the variability to be expected in applications of the system.

2.1.1.2 Flaw Variation

The character of the flaws in the structure being inspected will have a significant influence on the inspection outcome. There are two elements of flaw character that impact the demonstration: the physical characteristics of the specimens containing the flaws, and the physical properties of the flaws in the specimens. The inspection system will be designed to detect flaws of a defined size range at a location in a structural element defined at least by a material type and geometrical configuration combination. A fixed set of specimens containing flaws will be inspected, and these specimens either

must be of this combination or the assumption must be made that differences in inspection response in the specimens is identical to that obtained in the real application. (Although analytical methods are being sought for inferring POD(*a*) from different material/geometry configurations, no acceptable method for correlating between configurations is currently available.)

The flaws in the specimens must be as close as possible to the flaws that will be in the real structures and of sizes that span the region of interest for the POD(*a*) analysis. The assumption of equivalent response to the real inspection is implied when the results of the demonstration are implemented. Experience with the inspection will dictate the degree of acceptance of the assumption. For example, electrical discharge machined (EDM) notches are not good substitutes for eddy current inspections of surface fatigue cracks but may be the only possible choice for subsurface ultrasonic inspections.

Inspection capability is expressed in terms of flaw size, but not all flaws of the same "size" will produce the same magnitude of inspection response. In general, the specimens used in NDE reliability demonstrations are very expensive to obtain and characterize in terms of the sizes of the flaws in the specimens. Each set of specimens will be inspected multiple times if other factors are being considered in the demonstration. From a statistical viewpoint, this restriction on the experimental design limits the sample size to the number of flaws in the specimen set. Multiple independent inspections of the same crack provide information only about the detection probability of that flaw and do not provide any information about the variability of inspection responses between different flaws. Stated another way, *k* inspections on *n* flaws is not equivalent to inspections of $n \bullet k$ different flaws, even if the inspections are totally independent. The number and sizes of flaws will be further discussed in the next subsection.

2.1.1.3 Hardware Variability

Accounting for variability due to differences in inspection hardware must first be considered in terms of the scope of the capability evaluation. Each component of the inspection system can be expected to have some effect, albeit small, on inspection response. The combinations of particular components into subsystems and complete inspection stations can also be expected to influence the response. Since different stations might have different POD(*a*) capabilities, a general capability objective must be set. Each station can be characterized, each facility comprising many stations can be characterized, or many facilities can be characterized. Ideally, stations would be randomly sampled for the scope of the desired characterization and a weighted average of responses would be used to estimate the POD(*a*) function. On a practical level this is seldom done for ostensibly identical equipment. (Note that an analogous problem exists when accounting for the human factors which will be discussed later.) More commonly, capability demonstrations are performed on one station, and the assumption is made that the characterization would apply to all stations. The POD(*a*) differences between stations are assumed to be negligible. This approach has been used, for example, in characterizing the ENSIP/RFC inspections on the ECIS.

The concept of performing capability demonstrations on a single workstation is directed at a complete inspection station (however defined), but the variability of interchangeable components of a system can often be directly assessed. For example, experience has shown that different eddy current probes produce different responses when all other factors are constant. If a single probe is used to demonstrate the capability of an eddy current system, the estimated POD(*a*) function applies to the relevant inspections using that probe. However, if the POD characterization is to be used for in-service inspections using any such probe, an assumption is required that the probe is representative of the entire population. If a larger demonstration is affordable, the inspections could be performed using a random sample of probes from the available population. The analysis method must then account for the fact that multiple inspections of each crack were made with the different probes. The resulting characterization would better represent an inspection for a randomly selected probe.

Accounting for the variation from more than one source is more complex. Care must taken to ensure that the multiple sources are balanced in the analysis of the data and that the correct analysis procedures are used. For example, in the early evaluations of the ECIS for the ENSIP/RFC applications, there was considerable interest in the inherent variability in response from repeated, identical inspections and in the variability that results from different probes with their associated recalibration changes. (Other factors were initially considered but were later ignored after it was shown that they had no affect on POD(*a*) for the system.) The specimen sets would be inspected three times: twice with one probe and once with a second probe. The data from the three inspections, however, could not be combined in a single analysis since such an analysis would skew the results toward the probe with double representation. Thus, one analysis would be performed to estimate the inherent repeat variability and a second analysis would be performed to estimate the probe-to-probe variation. The results would then be combined to arrive at a POD(*a*) function that accounted for both sources of variation. It might be noted in this context that the repeat variability was negligible as compared to the variability that results from recalibration and probe changes. The demonstration plan was later modified by performing the third inspection with a third probe to better estimate the more significant between-probe variation.

Factorial type demonstrations are an efficient approach to simultaneously account for several significant factors. However, such demonstrations for more than a couple of factors require many inspections of the specimen set. More sophisticated statistical experimental designs might be employed, but the actual choice of such a design and the analysis of the data are driven by the specific objectives of a particular experiment. Discussion of such designs is beyond the scope of this report.

2.1.1.4 Human Factors

When inspectors play a significant role in the find/no find decision, they are an integral component of the NDE system. In such common inspection scenarios, human factors can contribute significantly to the variability in inspection results. In this context, human factors

refer to both the dynamic capabilities of individual inspectors and the user friendliness of the inspection tools in the environment of the application. Experiments have been conducted to quantify some of the environmental effects of human factors and data from some demonstration experiments have been interpreted in terms of the level of training and experience of the inspectors (see, for example, Spencer, et. al. [8] and Lewis, et. al. [9]). However, the effects and interactions of human factors on inspection results have not been characterized in the research. Rather, to the extent possible, NDE systems are automated to minimize the effect attributed to the inspector.

In a nonautomated inspection, many human factors potentially influence the inspection decision, and they cannot all be accounted for in a capability demonstration. At some level, the representative inspection assumption will be required. Given that the mechanical aspects of the NDE system and inspection environment are held constant, differences between inspectors, if ignored, can cause a biased capability characterization. Again, the objective of the capability characterization must be stated in advance. If each inspector is being evaluated, a separate POD(a) function for each is estimated. If a single POD(a) function is to be estimated for an entire facility, the inspectors in the demonstration must be randomly sampled in proportion to the percent of such inspections that each performs. Alternately, inspectors might be categorized by capability as implied by certification level, for example. A random sample of the inspectors from each level could be selected to arrive at a composite POD(a) for the level, and a weighted average would be calculated based on the percent of inspections performed by each level. An example of designing such a demonstration is given in Sproat, et. al. [10] and Hovey, et. al. [11]. Example results from the evaluation of a population of inspectors can also be found in Davis [12].

2.1.2 Sample Size Requirements

Sample sizes in NDE reliability experiments are driven more by the economics of specimen fabrication and flaw characterization than by the desired degree of precision in the estimate of the POD(a) function. POD(a) functions that appear reasonable can often be obtained from applying the maximum likelihood analysis to an inspection of relatively few specimens. Totally unacceptable results can also be obtained from inspecting specimens containing too few flaws or from inspection results that are not reasonably represented by the assumptions of the models. Therefore, it must be recognized that the confidence bound calculation is based on asymptotic (large sample) properties of the estimates and that there are minimal sample size requirements that must be met to provide a degree of reasonable assurance in the characterization of the capability of the system.

Larger sample sizes in NDE reliability experiments will, in general, provide greater precision in the estimate of the POD(a) function. However, the sample size is determined from the number of cracks in the experiment and there is an information content coupling with the flaw sizes that must also be considered. The effect of this coupling manifests itself differently for the $\hat{a}$ versus a and hit/miss analyses.

2.1.2.1 Sample Size Requirements for $\hat{a}$ versus a Analysis

When the flaw decision is made on the basis of a recorded response, $\hat{a}$, to the inspection stimulus, the data are known as $\hat{a}$ versus a inspection results. The added information from the $\hat{a}$ values provides a better approach to estimating POD(a). An example of $\hat{a}$ versus a data from a capability demonstration is presented in Figure 1. When the inspection response is greater than a preset detection threshold, a flaw is indicated for the site. In a capability demonstration, the minimum signal threshold is set as low as possible with respect to noise. Detection thresholds are later set that will yield a desired a_{90} value with an acceptable rate of extra indications. Extra indications are flaw indications at sites with no known flaws. Extra indications can be the result of noise or large responses from insignificant flaws. However, they can also result from anomalies that do not impair structural integrity.

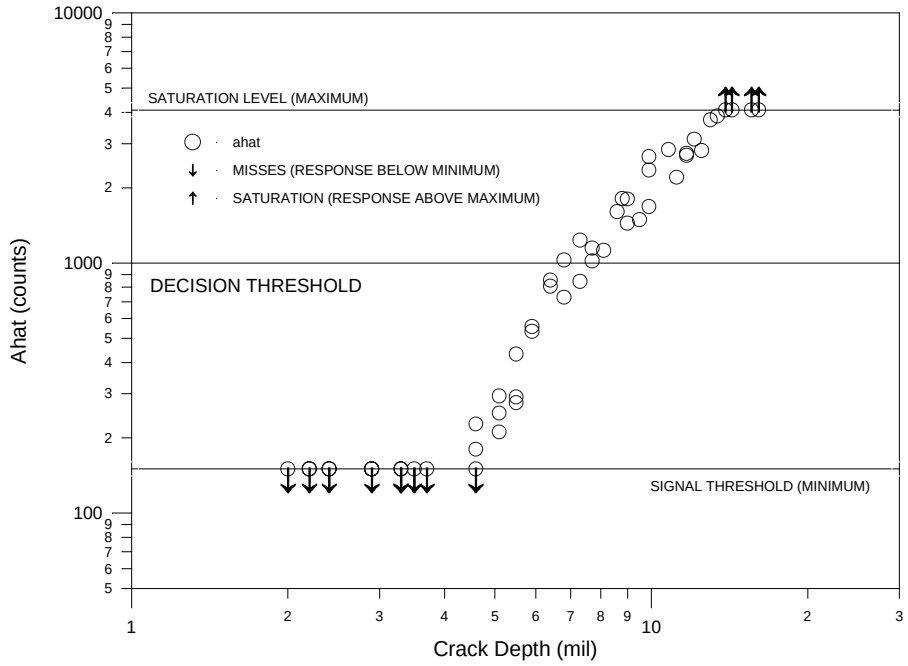


Figure 1. Example Plot of $\hat{a}$ versus a Data

The recorded signal response, $\hat{a}$, provides significantly more information for analysis than a simple flaw or no flaw decision of a hit/miss inspection response. The POD(a) model is derived from the correlation of the $\hat{a}$ versus a data, and the assumptions concerning the POD(a) model can be tested using the signal response data. Further, the pattern of $\hat{a}$ responses can indicate an acceptable range of extrapolation. Therefore, the range of crack sizes in the experiment is not as critical in an $\hat{a}$ versus a analysis as in a hit/miss analysis. For example, if the decision threshold in Figure 1 was set at 1000 counts, only the cracks with depths between about 6 and 10 mil would provide information that contributes to the estimate of the POD(a) function. The larger and smaller cracks are always found or missed and would have provided little information about the POD(a) function in a hit/miss analysis. In an $\hat{a}$ versus a analysis, however, all of the recorded $\hat{a}$ values provided full

information concerning the relation between signal response and crack size, and the censored values at the signal minimum and maximum limits provided partial information. The parameters of the POD(a) function are derived from the distribution of $\hat{a}$ values about the median response for flaws of size a. Assumptions necessary for characterizing this distribution are readily evaluated with the $\hat{a}$ versus a data.

Because of the added information in the $\hat{a}$ data, a valid characterization of the POD(a) function with confidence bounds can be obtained with fewer flaws than are required for the hit/miss analysis. It is recommended that at least 30 flaws, whose results can be recorded in $\hat{a}$ versus a form, be available for demonstrations. Increasing the number of flaws increases the precision of estimates, so the specimen test set should contain as many flawed sites as economically feasible. The analysis will provide parameter estimates for smaller sample sizes, but the adequacy of the asymptotic distributions of the estimates is not known.

2.1.2.2 Sample Size Requirements for Pass/Fail Analysis

In a hit/miss capability demonstration, the inspection results are expressed only in terms of whether or not the flaw of known size was detected. There are probabilities associated with each inspection outcome, and the analysis assumes that this probability increases with flaw size. Since it is assumed that the inspection process is in a state of control, there is a range of flaw sizes over which the POD(a) function is rising. In this flaw size range of inspection uncertainty, the inspection system has limited discriminating power in the sense that detecting or failing to detect would not be unusual. Such a range might be defined by the interval ($a_{0.10}$, $a_{0.90}$), where a_p denotes the flaw size that has probability of detection equal to p; that is,

$$\mathrm{POD}(a_p) = p \tag{1}$$

Flaws smaller than $a_{0.10}$ would then be expected to be missed, and flaws greater than $a_{0.90}$ would be expected to be detected.

In a hit/miss capability demonstration, flaws outside the range of uncertainty do not provide as much information concerning the POD(a) function as cracks within this range. Cracks in the almost certain detection range and almost certain miss range provide very little information concerning probability of detection. In the hit/miss demonstration, not all flaws convey the same amount of information and the “effective” sample size is not necessarily the total number of flaws in the experiment. For example, adding a large number of very large flaws does not increase the precision in the estimate of the parameters of the POD(a) function.

Ideally, all of the cracks in a hit/miss demonstration would have 80 percent of their sizes in the ($a_{0.10}$, $a_{0.90}$) range of the POD(a) function. However, it is not generally possible to have a set of specimens with such optimal sizes for all demonstrations. The demonstrations are being conducted to determine this unknown range of sizes for the NDE system being

evaluated. Further, because of the high cost of producing specimens, the same sets of specimens are often used in many different demonstrations. To minimize the chances of completely missing the crack size range of maximum information and to accommodate the multiple uses of specimens, the sizes of flaws in a specimen set should be uniformly distributed between the minimum and maximum of the sizes of potential interest. Mil-HDBK-1823 [4] recommends that a minimum of 60 flaws should be distributed in this range, but as many as are affordable should be used. This minimum sample size recommendation was the result of subjective considerations as to the number needed to make the asymptotic assumptions reasonable, experience in applying the model to data, and the results of analysis from a number of simulated POD demonstrations [13,14,15].

2.1.2.3 Unflawed Inspection Sites

In the context of the preceding discussion, sample size refers to the number of known flaws in the specimens to be inspected during the capability demonstration. The complete specimen set should also contain inspection sites that do not contain any known flaws. If the inspection results are of the hit/miss nature, at least twice as many unflawed sites as flawed sites are recommended. The unflawed sites are necessary to ensure that the NDE procedure is truly discriminating between flawed and unflawed sites and to provide an estimate of the false call rate. If the NDE system is based on a totally automated $\hat{a}$ versus a decision process, many fewer unflawed sites will be required. If any $\hat{a}$ values are recorded at the unflawed sites, their magnitude would provide an indication of the minimum thresholds that might be implemented in the application.

2.1.3 Specimen Flaw Size Requirements

As noted in the previous subsection, it is necessary that the specimens used in a demonstration have inspection sites that contain flaws with sizes in the range of interest. Inspection capability evaluations have been conducted with real components that were later destructively inspected to characterize the flaws that were in the structure, Lewis, et al. [9]. This after-the-fact determination of flaw sizes is not cost-effective for widespread determination of POD(a) functions. The more common practice is to design a specimen that has representative material and geometric properties and introduce flaws of the appropriate sizes. This is the approach used in Mil-HDBK-1823 [4] and has been used in all evaluations of the RFC/ENSIP ECIS.

In the context of a general flaw, there are a number of approaches to introducing flaws in a set of specimens. Examples of such approaches are EDM notches to simulate subsurface defects, artificially induced corrosive thinning, and fatigue-induced cracks. The key issue with an artificial flaw is the degree to which the flaw is representative, an issue that is beyond the scope of this report. The key issue with induced fatigue cracks is the characterization of the size of the crack. Because crack size directly impacts the estimate of the POD(a) function, the effects of errors in the determination of the specimen crack sizes were investigated and are discussed in the following.

Cracks are introduced in POD specimens by fatigue cycling the specimen until a target crack length is obtained. If necessary, a starter notch is introduced prior to cycling and all remnants of the starter notch are later removed. At present, there are no nondestructive methods for measuring the crack depth or determining crack shape. Damage tolerance life calculations are often driven by crack depth. In this circumstance, POD(a) is either characterized in terms of depth, or a relation is assumed that correlates the size used in the analysis with the depth of the fracture mechanics calculations. For visual inspections, the observable length determines the inspection response and the appropriate measure of size for the POD(a) characterization. The depth characterization, if needed, and the uncertainty in the measure of crack length are not issues for the relatively large target sizes of cracks of visual inspections. However, for eddy current inspections, the response is a function of crack area (i.e., length and depth). A better estimate of POD(a) can be obtained if the crack depth is also accounted for in the estimate of crack size.

In some applications, fatigue cracks are assumed to have a fixed aspect ratio of length to depth with a semicircular shape. The size measurement used to estimate POD(a) is calculated by assuming an equivalent area for the measured crack length and estimated depth. This approach is also used when both length and depth measurements are possible, as, for example, from a corner crack at a bolt hole. In other applications, the crack depth is estimated by correlating length and depth measurements from a destructively inspected subset of the specimen set.

The mis-sizing of cracks affects the estimation of POD(a) because some of the apparent scatter in the inspection response may be due to errors in the crack size. Such potential problems are best portrayed by an example. The following example is based on $\hat{a}$ versus a data from the RFC/ENSIP ECIS, but the concepts would apply to any inspection system whose response depends on more than one flaw dimension.

Figure 2 presents a set of $\hat{a}$ versus a data obtained on IN100 flat plate specimens using an RFC/ENSIP ECIS. Twenty one of the cracks have estimated depths less than 4 mil. The inspection response for 9 of these small cracks was greater than the minimum (signal) threshold $\hat{a}$ value of 100 counts. The pattern and magnitude of the $\hat{a}$ responses for these small cracks are significantly different from those of the larger cracks. Are the small cracks incorrectly sized or are the small cracks responding differently to the eddy current stimulus than the larger cracks? While the response to small cracks may well be different from that of the larger cracks, there are also strong indications that the small cracks may be incorrectly sized.

Figure 3 presents a plot of length versus depth for the specimens that were destructively inspected to establish the correlation between length and depth for this specimen set. A straight line was fit to the points using least squares (regression). Crack lengths in the POD specimen set were measured using replications, and a depth was estimated for each crack as calculated from the regression equation. All cracks of the same length were assigned the same depth. More specifically, the differences in depths of individual cracks from the average for cracks of fixed length were ignored. However, the potential depth errors can be

quantified by prediction limits derived from the regression analysis. The regression line for predicting depth from length and the 50, 75, and 95 percent prediction limits are also shown in Figure 3. For a fixed crack length, 50, 75, and 95 percent of all cracks would be expected to have depths between the 50, 75, and 95 percent prediction limits, respectively.

This method of estimating crack depth assumes that the relationship between crack length and depth is well represented by the destructively inspected specimens. If the line were forced through the origin, the slope would define the aspect ratio. The least squares (regression) line of this example was not forced through the origin. There is a 1 mil negative bias that is negligible at longer lengths but has a significant relative impact at the smallest crack lengths. Because the crack starter notch is removed from the specimens, the crack aspect ratio is different for the smaller cracks, and the true relationship may not be linear at the small crack sizes. Thus, though particular data sets are reasonably represented by a straight line, a nonlinear length-to-depth relationship might be expected at the smallest crack sizes.

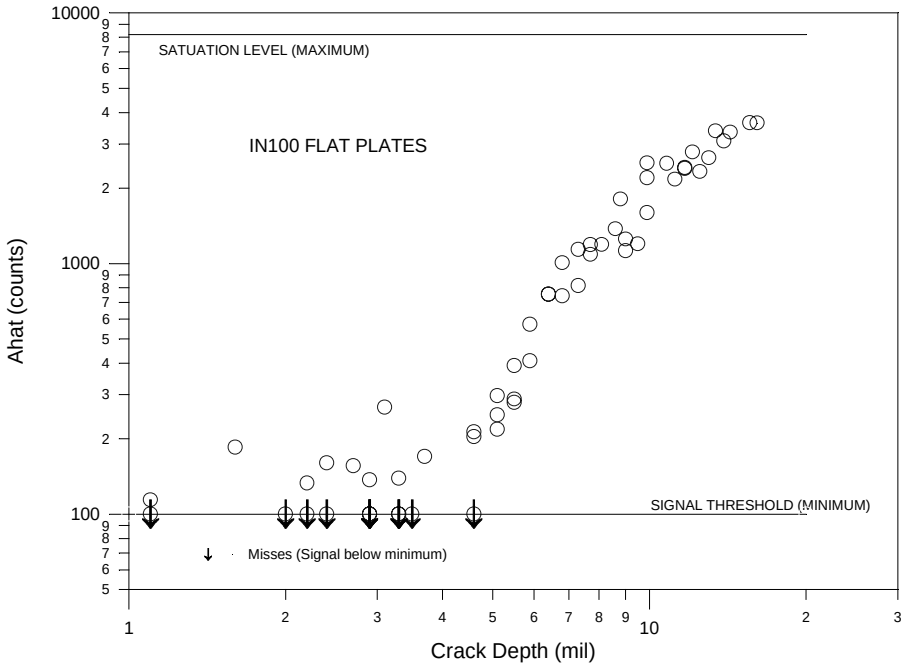


Figure 2. Example $\hat{a}$ versus a Data for Small Crack Depths

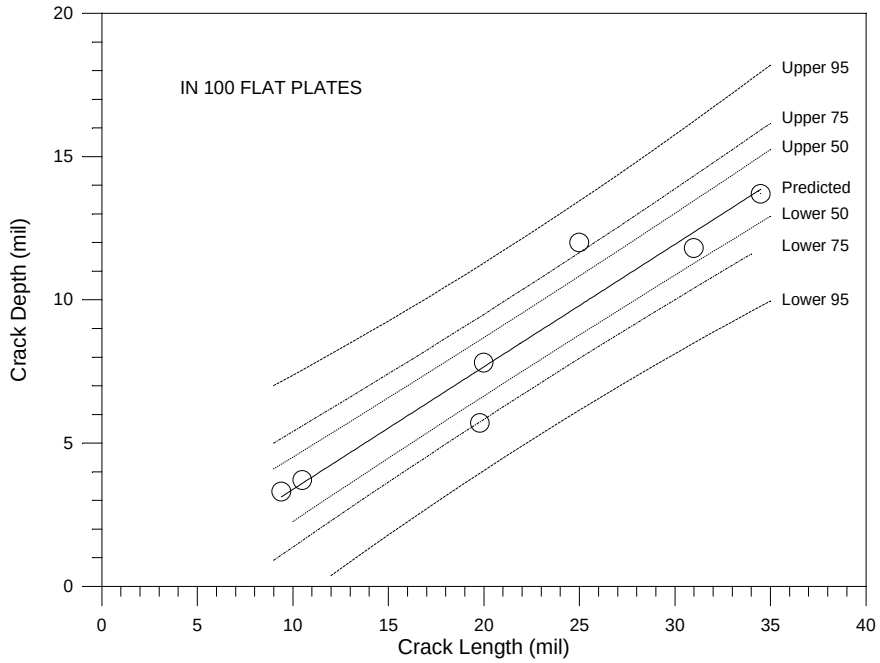


Figure 3. Correlation of Length with Depth for IN 100 Flat Plate Specimens

The crack depths for the erratic $\hat{a}$ versus a responses shown in Figure 2 were obtained by extrapolating the length-to-depth data well below the smallest crack in the destructively inspected specimens. The crack lengths represented in the destructively inspected cracks ranged from 9.4 to 34.5 mil. The crack lengths in the POD specimen set ranged from 3 to 39 mil with 14 cracks having lengths less than 9.4 mil. Two cracks in the POD specimen set were reported at a length of 5 mil for which the estimated depth was 1.1 mil.

The prediction intervals of Figure 3 indicate that half of the crack depths would be in error by more than 1 mil, 25 percent would be in error by more than 1.8 mil, and 5 percent would be in error by more than 3.6 mil. Possible sizing errors of these magnitudes would have a significant impact on the pattern of $\hat{a}$ versus a at the small sizes and this impact is accentuated when the usual log transformation is used. For example, in Figure 2, a random 1 or 2 mil sizing error in crack depths greater than 10 mils would have an insignificant effect in the POD analysis. However, a 1 or 2 mil sizing error in cracks with depths of 2 or 3 mil would have a large effect in the POD analysis. To demonstrate the range of potential mis-sizing that could result from this procedure for estimating depths, Figure 4 superimposes the 75-percent prediction limits for the crack depths of 2 and 3 mil on the $\hat{a}$ versus a data of Figure 2. Since one out of four cracks would be expected to have depths outside these limits, a 1 to 2 mill error in depth would be common. A 1 or 2 mil increase for the small cracks with $\hat{a}$ values greater than threshold would make them agree much better with the pattern from the larger cracks sizes. Errors of a couple of mil would be common and could produce the apparently aberrant behavior in $\hat{a}$ versus a response of Figure 2. Such behavior has been observed in many other data sets at small crack sizes.

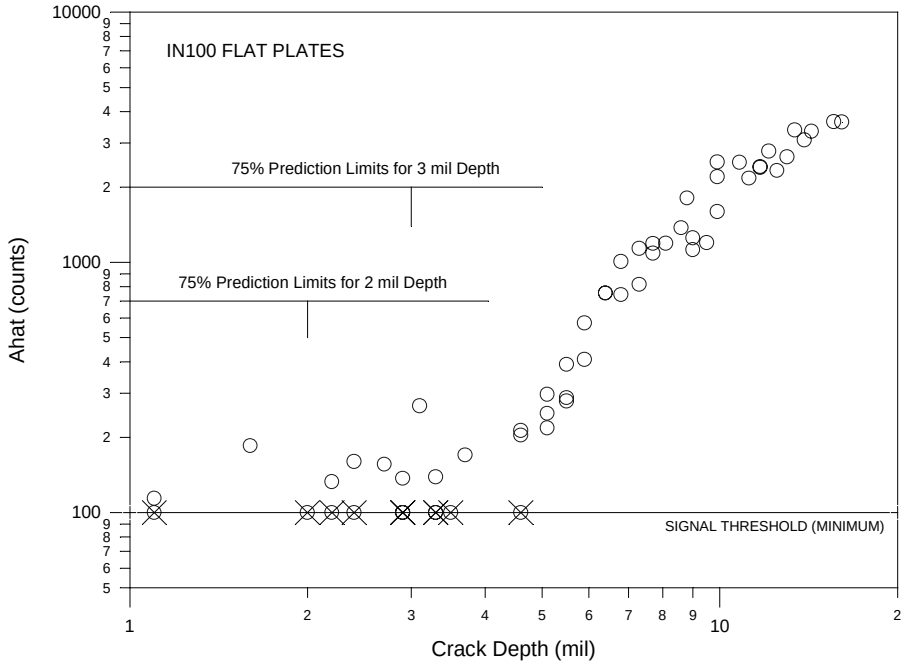


Figure 4. 75-Percent Limits on Crack Depths for IN 100 Flat Plate Specimens

The sizes of five of the cracks of this specimen set have been reevaluated. The specimen with an estimated depth of 1.6 mil was destructively inspected and found to have a measured depth of 4 mil. The other four, with depths originally estimated between 2.7 and 3.3 mil, were found to have larger lengths, and the revised estimated depths are between 3.5 and 4.8 mils. In this case, at least some of the erratic behavior in the original $\hat{a}$ versus a plot was definitely attributable to mis-sizing cracks in the specimen set.

For an NDE system in which the inspection response is proportional to the flaw area, a random, but unbiased, mis-sizing of cracks will lead to a conservative estimate of the POD(a) function. The inspection response, $\hat{a}$, to a flaw is determined by the true flaw geometry, regardless of the size that is assigned to it. If the sizing error is unbiased, the average response as a function of the size will be correct, but the errors in the size direction will increase the apparent scatter about the mean response. The increased scatter will produce a flatter, somewhat more conservative POD(a) function – i.e., higher a_{90} values. Statistical methods could be used to account for the increased scatter, but such methods require external measures of the standard deviation of the sizing errors. At present, the sizing errors can be measured only by destructive tests of the fatigue-cracked specimens. Since the degree of conservatism in POD(a) characterizations has been acceptable, the cost of a sufficiently large sample of destructively inspected specimens to obtain a reasonably precise estimate of their standard deviation is not warranted.

Errors in the crack sizes of the POD specimens are invisible in a find/no find inspection. While different find/no find decisions for two cracks that have been assigned the same

size may be due to one crack being smaller than the other, there is nothing in the data to indicate this cause for the differing decisions. In $\hat{a}$ versus a data, an erratic pattern of responses at very small sizes might be the result of the inspection system, but might also be the result of mis-sizing the cracks in the specimens. When the pattern of the responses is not compatible with assumptions required by the $\hat{a}$ vs a analysis, the data have been partitioned and separate analyses have been applied to the partitions. No concessions have been made to the possible sizing errors. However, in the evaluation of the RFC/ENSIP ECIS, it was quite common to exclude the smallest cracks from an analysis because of the erratic $\hat{a}$ responses. This exclusion has been possible because there were other cracks with responses that were less than the minimum recordable level (signal threshold), and because achievable a_{90} values were above the cutoff size. The pattern of $\hat{a}$ versus a for these small cracks was irrelevant.

Considerable care should be taken in the estimation of the sizes of the flaws in the POD specimens. This intended care is directed at both the specific measurements of the physical property (e.g., crack length) and at the method for calculating the POD size metric (e.g., crack depth from regression analysis) from the measurements. The effects of sizing errors are acceptably conservative for larger flaws. The sizing of small flaws is difficult because a) the measurement of very small flaws is inherently more difficult, b) aspect ratios that have been observed in fatigue cracks may not apply to small cracks in specimens, and c) the prediction of depth from length produces proportionally larger errors for small than for large flaws. Since sizing errors for very small flaws can significantly influence the apparent inspection results, extra attention should be given to the small flaws in the POD(a) analysis. Anomalies in $\hat{a}$ versus a behavior in the small flaw regime are more likely to be attributable to size errors than to inspection response.

2.2 $\hat{a}$ Versus a Analysis

All NDE systems make find/no find decisions by interpreting the response to an inspection excitation. In some inspections, the response is a recordable metric, $\hat{a}$, that is related to the flaw size. Find/no find decisions are made by comparing the magnitude of $\hat{a}$ to the decision threshold value, $\hat{a}_{dec}$. The $\hat{a}$ versus flaw size analysis is a method of estimating the POD(a) function based on the correlation between $\hat{a}$ and flaws of known size, a. The general formulation of the $\hat{a}$ versus a model is expressed as

$$\hat{a} = f(a) + \delta, \qquad (2)$$

where f(a) represents the average (or median) response to a crack of size a, and δ represents the sum of all the random effects that make the inspection of a particular size a crack different from the average of all cracks of size a. In principle, any f(a) and distribution of δ that fit the observations can be used. However, if f(a) is linear in a and δ is normally distributed with constant standard deviation, σ_δ, then the resulting POD(a) function is a cumulative normal distribution function. (Monotonic transformations of $\hat{a}$ or a can also be analyzed in this framework.) This specific formulation of the $\hat{a}$ versus a relation has fit, or been adaptable to, the data from many capability demonstrations and is the focus of this report. In particular, all RFC/ENSIP ECIS capability evaluations have been made in

terms of a linear fit to ln $\hat{a}$ versus ln a. The computer code specified in Mil-HDBK-1823 [4] is based specifically on this linear ln $\hat{a}$ versus ln a model.

This subsection briefly reviews the $\hat{a}$ versus a analysis, summarizes the RFC/ENSIP ECIS experience in terms of parameter variation and POD(a) sensitivity to the residual scatter σ_δ, and details the use of transformations that can be implemented using the updated POD(a) program. The estimates of model parameters and their sampling distributions are based on maximum likelihood analysis as described in Mil-HDBK-1823 [4] and Berens [5]. The details of these calculations will not be repeated here.

2.2.1 $\hat{a}$ versus a Model Formulation

The formulation of the $\hat{a}$ versus a analysis that has been used exclusively in the evaluation of the RFC/ENSIP ECIS is expressed in terms of the natural logarithms of $\hat{a}$ and a:

$$\ln \hat{a} = B_0 + B_1 \bullet \ln a + \delta, \tag{3}$$

where δ is normal (0, σ_δ). For a decision threshold of $\hat{a}_{dec}$, the following applies:

$$\text{POD}(a) = \Phi\,[(\ln a - \mu)/\sigma], \tag{4}$$

where $\Phi(\bullet)$ is the cumulative standard normal distribution function and

$$\mu = (\ln \hat{a}_{dec} - B_0) / B_1 \tag{5}$$

and

$$\sigma = \sigma_\delta / B_1. \tag{6}$$

The calculation is illustrated in Figure 5. The parameters of the $\underline{\hat{a}}$ versus a model (B_0, B_1, and σ_δ) are estimated from the data of the demonstration specimens. The probability density function of the ln $\hat{a}$ values for a 13 mil crack depth is illustrated in the figure. The decision threshold in the example is set at $\hat{a}_{dec}$ = 165. The POD for a randomly selected 13 mil crack would be the proportion of all 13 mil cracks that would have an $\hat{a}$ value greater than 165, i.e., the area under the curve above 165. In this example, the decision threshold was selected so that POD(13) = 0.90.

The characterization parameter of major interest in the completely automated inspections of the RFC/ENSIP ECIS is a_{90}, the crack length for which POD(a_{90}) = 0.90. In this formulation of the $\hat{a}$ versus a analysis, the following equation applies:

$$a_{90} = \exp(\mu + 1.282\,\sigma) \tag{7}$$

The value a_{90} is completely determined from the fit to the mean response as a function of crack size as determined by B_0 and B_1 and the scatter about the mean as determined by σ_δ.

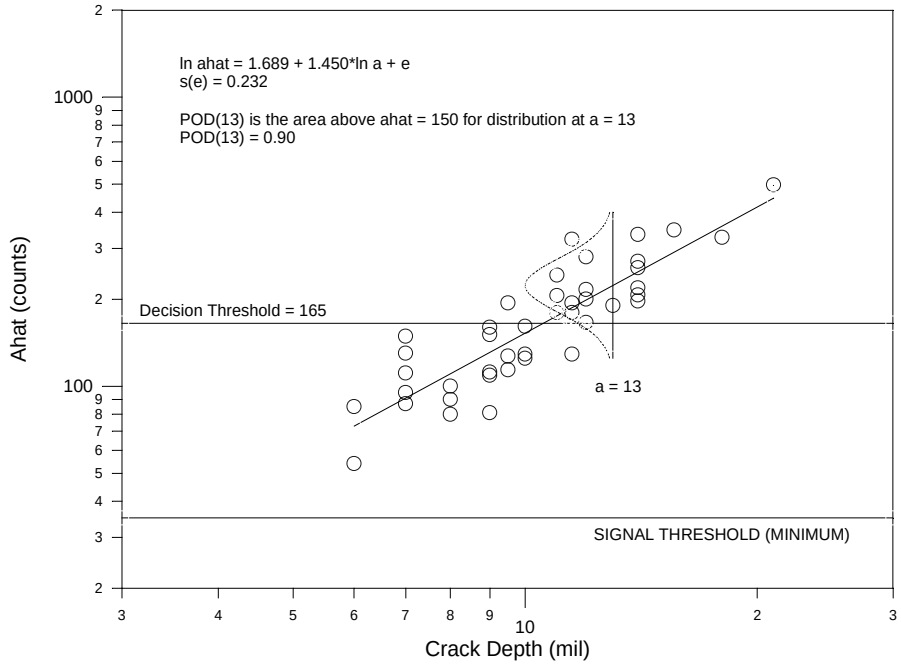


Figure 5. Example POD(*a*) Calculation from $\hat{a}$ versus *a* Data

The preceding formulation of the $\hat{a}$ versus *a* model is based on three assumptions: a) the mean of the log responses, ln $\hat{a}$, is linearly related to log crack size, ln a; b) the differences of individual ln $\hat{a}$ values from the mean response have a normal distribution; and c) the standard deviation of the residuals, σ_δ, is constant for all a. These assumptions can be tested using the results of the data from the demonstration. When the assumptions are not acceptable, current practice is to restrict the analysis to a range of crack sizes for which the assumptions are acceptable.

When applying the model in the evaluations of the RFC/ENSIP ECIS, the assumptions of equal standard deviation of residuals and linearity of ln $\hat{a}$ versus ln *a* have been rejected on occasion. When these assumptions were not justified, the data would be segmented into crack size ranges for which the assumptions were acceptable. On occasion, the number of cracks in some analyses was below the recommended sample size, and discrete jumps at the end of the intervals were a source of ambiguity. In all cases, a conservative answer was reported.

Unequal residual standard deviations of the ln $\hat{a}$ residuals were usually caused by an increase in scatter at small flaw sizes as previously discussed in subsection 2.1.3. The effect of ignoring this increase will be investigated in terms of an error analysis of the a_{90} values. Nonlinearity was typically caused by a characteristic concave downward trend in $\hat{a}$ values as the flaw size increased. The cause of this response is attributed to the probe size. The use of transformations to mitigate this problem will be presented. The following subparagraphs address these issues.

2.2.2 Variability of $\hat{a}$ About Mean Response

The a_{90} values are explicit functions of the parameters of the fit of $\hat{a}$ versus a data from NDE capability demonstrations. Because of the increase in the scatter about the fit due to the potential mis-sizing of the cracks in the demonstration, it is instructive to investigate the sensitivity of a_{90} to the estimate of the residual standard deviation. In all of the demonstrations of the RFC/ENSIP ECIS, there were repeat inspections of the same sets of specimens under "constant" conditions. The inherent variability of the a_{90} values from these nominally identical inspections provides a baseline for judging potential scatter in capability demonstrations. First, a summary of the uncertainty in results from nominally identical inspections RFC/ENSIP ECIS will be presented. The results of a sensitivity analysis on the $\hat{a}$ versus a parameters will be interpreted in the light of this baseline.

2.2.2.1 Inherent Variability in RFC/ENSIP ECIS Demonstration Data

Over 100 POD(a) capability evaluations have been conducted on the RFC/ENSIP ECIS. The objective of these evaluations was to characterize the relation between a_{90} and the decision threshold, $\hat{a}_{dec}$, for combinations of materials and geometry. (A few evaluations were conducted to look at differences in inspection stations, but no concerted efforts were made to characterize individual stations or to account for variation in results due to different stations.) When the demonstrations of the ECIS started, about 1990, three inspections would be performed for each specimen set. Two inspections were performed with one probe to determine the minimum degree of repeatability including the variation due to recalibration. An additional inspection was performed with a second probe of identical type to measure the probe-to-probe variation as confounded with the variation due to calibration. It was noted that the variability due to repeat inspections with the same probe was negligibly small, and this inspection was later replaced by one using a third probe.

Figure 6 presents an example of $\hat{a}$ versus a response from three ECIS inspections using three different probes on the same titanium bolt hole specimen set. The three $\hat{a}$ values from the same crack tend to cluster about the mean for that crack with the differences being due to probe differences, calibration precision, and all of the other factors that cannot be controlled between the inspections. In this example, the Probe 1 recordings tend to be larger than the Probe 2 and 3 recordings. However, there is far more dispersion between cracks of the same size than in the three readings of the same crack.

Because of the differences in $\hat{a}$ values from the multiple inspections, different estimates of a_{90} resulted from the analysis of the data from the individual probes. The RFC/ENSIP ECIS evaluations provide a baseline for the degree of scatter in the estimated a_{90} values which are the result of only such natural sources of variability. Note that in the analysis of the ECIS data, care was always taken to ensure that the linearity, equal variance, and normality assumptions were reasonable for the reported crack size range of validity. The scatter in a_{90} values in these data sets is not due to deviations from model assumptions. On 55 of the bolt hole and flat plate evaluations, an $\hat{a}$ versus a analysis was performed for

each individual probe or repeat inspection, and a_{90} values were calculated for a common, relatively low decision threshold. A low decision threshold was selected because the a_{90} differences are magnified at higher thresholds.

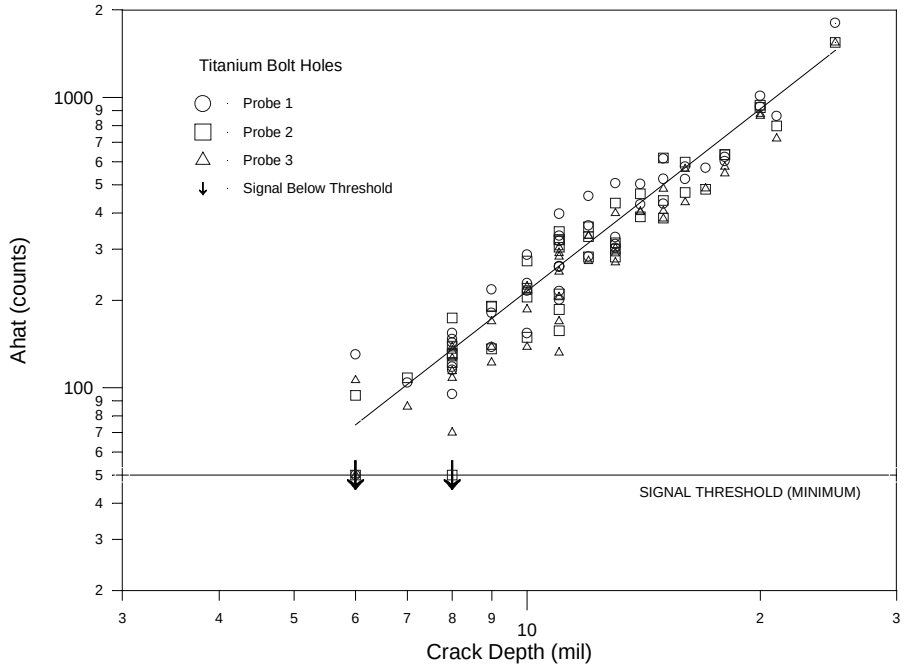


Figure 6. Example $\hat{a}$ versus a Inspection Data with Three Probes

The standard deviation of a_{90} values for each of the 55 data sets and pooled estimates for geometry by material combinations were calculated. Eighteen percent of the standard deviations were greater than 1 mil and it was inferred that probe and recalibration differences can easily result in a 2 mil difference in a_{90}. Table 1 presents the composite standard deviations for the four combinations of titanium and nickel base flat plates and bolt holes. There is no consistent pattern between materials and geometry but, the nickel base flat plates displayed the most variation in the a_{90} estimates.

Table 1. Standard Deviation of a_{90} Values from Repeat Inspections under Identical Conditions for Geometry by Material Combinations

	Flat Plates	Bolt Holes
Titanium	0.862 mil	1.004 mil
Nickel Base	1.221 mil	0.295 mil

The variability in the $\hat{a}$ recordings of the ECIS can be further characterized by introducing a random term for between-probe recalibration sources of variation. Let $\hat{a}_{ij}$ represent the j^{th} inspection response to crack i. Then the following equation applies:

$$\hat{a}_{ij} = B_0 + B_1 \ln a_i + c_i + p_{j(i)}, \qquad (8)$$

where c_i is the difference in average response of crack a_i from the average of all cracks of size a_i. The $B_0 + B_1 \ln a_i$ is the average response of all cracks of size a_i. The random term $p_{j(i)}$ is the difference in response from probe j and recalibration from the average of all potential inspections of crack a_i using a population of probes.

In this formulation of the $\hat{a}$ versus a model, δ of equation (3) is given by the following:

$$\delta_i^* = c_i + p_{j(i)}, \tag{9}$$

Because the fit of the ln $\hat{a}$ data is made in terms of average $\hat{a}$ values for each crack, σ_δ of equation 6 is estimated from the following:

$$\sigma_\delta^2 = \sigma_{\delta^*}^2 + (k\text{-}1)\ \sigma_p^2\ /k, \tag{10}$$

where k is the number of different probes. The two components of the total variability in equation 10 are estimated from the inspection results. The $\sigma_{\delta^*}^2$ is the residual variance from the regression analysis of the possibly censored data. The σ_p^2 is estimated by pooling the variances from the multiple $\hat{a}$ values obtained from the k probes on each crack. See Berens [5] for details.

Estimates of the ECIS components of variation expressed in equation 10 provide a baseline for the degree of uncontrolled scatter that can be expected in evaluations of automated eddy current systems. Table 2 presents median standard deviations for geometry by material combinations. The Repeat medians are from those evaluations for which a repeat inspection with the same probe was conducted. The repeat standard deviation reflects the minimum scatter that is attainable with recalibration. Also included in Table 2 is the median σ of the POD(a) equation.

Table 2. Median Standard Deviations of ln $\hat{a}$ by Source of Variability for Material by Geometry Combinations

	Flat Plates		Bolt Holes		Other Geometries	
	Titanium	Nickel	Titanium	Nickel	Titanium	Nickel
Cracks, σ_{δ^*}	0.332	0.255	0.295	0.399	0.321	0.259
Probes, σ_p	0.106	0.112	0.117	0.103	0.175	0.115
Repeats, σ_r	0.071	0.063	0.063	0.066		
σ	0.240	0.152	0.164	0.228	0.221	0.166

Since, as indicated by equation 10, σ_δ is calculated from a sum of the squares of the standard deviations of sources of variability, the standard deviations of the crack to crack residuals dominate in the calculation of σ. For example, the between-probes standard deviation is largest relative to the between-cracks standard deviation for the titanium, other geometries. Not including the probe-to-probe variability in the estimate of the median σ would change the estimate by only 2.5 percent.

The steepness of the POD(a) function is determined by σ. Table 3 presents the range and quartile estimates of σ that were obtained from the ECIS evaluations for the material by geometry combinations. It might be noted that the larger values of σ were obtained in some of the earlier evaluations.

Table 3. Statistical Summary of σ Values Obtained in ECIS Evaluations

	Flat Plates		Bolt Holes		Other Geometries	
	Titanium	Nickel	Titanium	Nickel	Titanium	Nickel
Minimum	0.075	0.050	0.111	0.177	0.090	0.101
1st Quartile	0.131	0.083	0.148	0.207	0.168	0.110
Median	0.240	0.152	0.164	0.228	0.221	0.166
2nd Quartile	0.408	0.300	0.215	0.280	0.321	0.262
Maximum	0.636	0.706	0.285	0.516	0.507	0.345

2.2.2.2 Sensitivity of a_{90} to Parameter Variations

The sensitivity of a_{90} to changes in the scatter of the $\hat{a}$ residuals from the fit obtained from the ln $\hat{a}$ versus ln a model can be evaluated in terms of a_{90} ratios. The following analysis provides a baseline for judging the effects of a nonconstant residual standard deviation in the ln $\hat{a}$ versus ln a analysis. It might be noted that the B_0 and B_1 terms are somewhat set by the design and calibration of the system. Because the analyses are performed in terms of the logarithm of the response and calibrations are performed in decibels, calibration errors are introduced as additive errors to B_0 and are reflected in the estimate of the probe-to-probe variability.

Let μ and σ be the true POD(a) parameters for an inspection system and let $\Delta\mu$ and $\Delta\sigma$ represent the errors in the estimates of the parameters. The percent error in the a_{90} value obtained from the perturbed parameters can be calculated from equation 7as follows:

$$\text{Percent Error} = 100\,[\exp(\Delta\mu)\exp(1.282\,\Delta\sigma) - 1] \qquad (11)$$

The $\Delta\mu$ depends on possible errors in B_0 and B_1, and $\Delta\sigma$ depends on possible errors in σ_δ and B_1. Assuming no errors in B_0 and B_1, the percent error in a_{90} depends only on the error in the estimate of the standard deviation of the residuals of the $\hat{a}$ values from the fit, σ_δ. That is,

$$\Delta\sigma = \Delta\sigma_\delta / B_1 \qquad (12)$$

and

$$\text{Percent Error} = 100\,[\exp(1.282\,\Delta\delta_\delta / B_1) - 1]. \qquad (13)$$

Figure 7 presents the percent error in a_{90} as a function of potential errors in σ_δ for values of B_1 equal to l.5, 2.0, and 2.5. The average B_1 value in the ECIS evaluations was 1.8, and about 80 percent of the B_1 values were greater than 1.5.

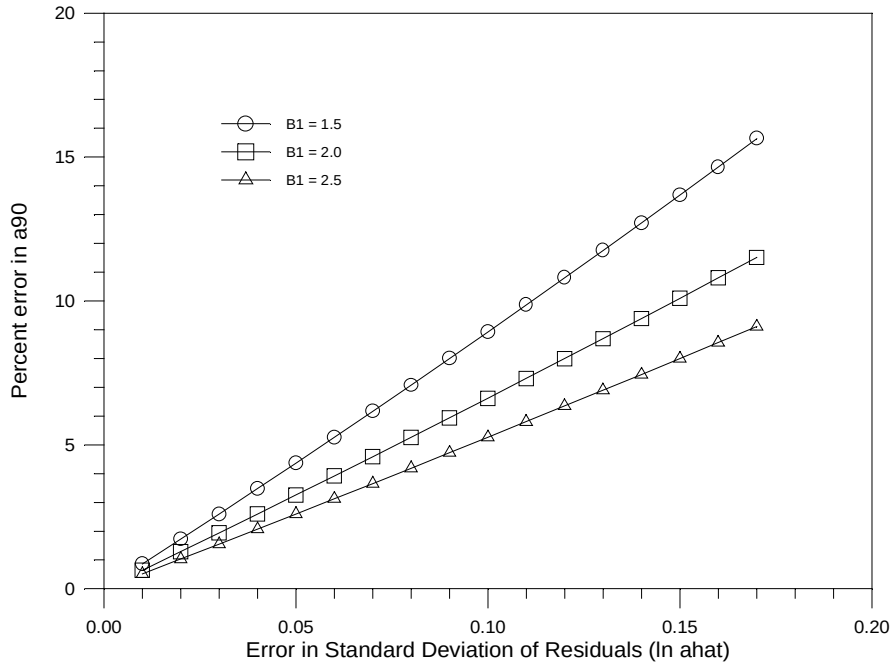


Figure 7. Percent Error in a_{90} versus Error in σ for Selected B_0

Table 2 indicates that the median values of σ_δ in the ECIS evaluation ranged from 0.25 to 0.4. An error of 0.10 in the estimate of σ_δ would represent a large relative error in this parameter but would typically produce less than a 10 percent error in the estimate of a_{90}. For target values of a_{90} less than 20 mil, this 1 to 2 mil magnitude of error is equivalent to that which has been observed in results from individual probe changes of otherwise identical inspections.

The equal variance of residuals assumption is evaluated using the Bartlett test for homogeneity of variance on a partition defined by large versus small cracks. In the ECIS evaluations, this assumption was usually rejected because the standard deviation at the smaller crack sizes was larger. When all of the cracks are in the same analysis, the composite standard deviation of the residuals is about the average of those for the two crack size ranges. It is possible that the Bartlett test could reject the assumption of equal residual standard deviations while the effect of ignoring the difference would be practically insignificant. For example, if σ_δ for the two crack size ranges are within 0.1, then both ranges would be within about 0.05 of the composite standard deviation. If B_1 >1.0, Figure 7 indicates that the percent error in the estimate of a_{90} would then be less than 5 percent. If the target a_{90} value is 20 mils, the error in a_{90} would be less than 1 mil and this small error would be in the conservative direction. If the target a_{90} value is, say, 8 mils, the error would be less than 0.4 mils, but possibly in the non-conservative direction. Therefore, when the equal variance assumption is rejected, a careful investigation of the data may indicate that the analysis would still be acceptable. This is a particular concern in view of the previous discussion concerning the possibility that the larger standard deviation of residuals at the small crack sizes may well be due to errors in the sizing of the cracks. If mis-sizing of cracks is the cause of increased scatter in the ln $\hat{a}$ values about

the fit, the composite standard deviation will be larger and a more conservative estimate of a_{90} will be calculated.

Therefore, when the equal variance assumption is rejected, a careful investigation of the data may indicate that the analysis would still be acceptable. This is a particular concern in view of the previous discussion concerning the possibility that the larger standard deviation of residuals at the small crack sizes may well be due to errors in the sizing of the cracks. If mis-sizing of cracks is the cause of increased scatter in the ln $\hat{a}$ values about the fit, the composite standard deviation will be larger and a more conservative estimate of a_{90} will be calculated.

2.2.3 Transformation of Signal Response and Crack Size

The basic concept of the $\hat{a}$ versus a analysis is that POD(a) is determined from the distribution of the $\hat{a}$ responses about the function that relates mean response to crack size. The $\hat{a}$ versus a approach to the POD(a) analysis as implemented through equations 3 through 7 depends on a linear ln $\hat{a}$ versus ln a relation and a normal distribution of the ln $\hat{a}$ values about the mean. For three reasons, this particular $\hat{a}$ versus a model was originally programmed using the log linear, cumulative lognormal approach. First, the cumulative lognormal POD(a) model had previously been selected as an acceptable POD(a) model in an analysis of find/no find NDE reliability data from a completely different application [13]. Second, the assumptions required by the log linear model could be tested and were shown to fit a large number of data sets. Third, when the assumptions were not reasonably acceptable, the data could usually be partitioned into subsets for which the linearity assumptions were acceptable. However, other analysis approaches can be formulated and implemented. Because the log linear relation between $\hat{a}$ and a was often rejected in the analysis of RFC/ENSIP ECIS data an approach that is not completely dependent on this assumption is desirable.

There are two general approaches to providing for a nonlinear inspection response to crack size. These are transformations of $\hat{a}$ and/or a and the use of nonlinear functions to relate mean response to crack size. It is not feasible to preprogram an analysis that would be capable of acceptably fitting all future data sets with confidence bounds on a_{90}. However, it is feasible to provide for different transformations from which the data analyst can select an acceptable model for each specific application. Therefore, this report will focus only on the transformation approach. Box, et al. [16] and Neter, et al. [17] have extensive discussions on the use of transformations in regression analyses.

Note that the standard POD(a) model formulation is already being performed using the log transformation on both the crack size and the signal response. In the standard model formulation of equations 2 through 7, the analysis is expressed in terms of log transformations of both $\hat{a}$ and a. These log transformations have been hard programmed in the computer code that performs the analysis and are invisible to the user. However, there is nothing in the current application of the model that depends on this particular transformation.

The general formulation of the analysis that leads to the cumulative normal distribution equation for the POD(a) function can be expressed as follows:

If $$Y = B_0 + B_1 X + \delta \quad (14)$$

and δ is normally distributed with zero mean and a standard deviation of σ_δ, then

$$P(Y > Y_{th}) = \Phi[(Y-\mu_X) / \sigma_X], \quad (15)$$

where $\Phi(\bullet)$ is the cumulative standard normal distribution function and

$$\mu_X = (Y_{dec} - B_0) / B_1 \quad (16)$$

and $$\sigma_X = \sigma_\delta / B_1. \quad (17)$$

Equations 16 and 17 are similar to equations 3 through 6, the difference being that $Y = \ln \hat{a}$ and $X = \ln a$. Equation 18 is the inverse log transformation to convert to the correct units.

$$a_{90} = \exp(X_{90}) = \exp(\mu_X + 1.282\ \sigma_X) \quad (18)$$

Any monotonic transformations of a and $\hat{a}$ can be used. In general, let

$$Y = g(\hat{a}) \quad \text{or} \quad \hat{a} = g^{-1}(Y) \quad (19)$$

$$X = h(a) \quad \text{or} \quad a = h^{-1}(X) \quad (20)$$

and assume that the linear model of Equation 14 is reasonably acceptable for the range of application. Then $Y_{dec} = g(\hat{a}_{dec})$ and

$$a_{90} = h^{-1}(X_{90}) = h^{-1} (\mu + 1.282\ \sigma). \quad (21)$$

The intent of the transformations is to obtain both a linear mean $g(\hat{a})$ versus $h(a)$ relation and homogeneity of variance. (The standard deviation of differences from the straight line mean is independent of transformed crack size.) Transforming the crack size, a, will change only the shape of the mean response, not the distribution of residuals. Transforming $\hat{a}$ will change both the distribution of the residuals and the shape of the mean response. Three types of transformations have been found useful in meeting the linear regression assumptions by providing varying degrees of compression of the scale of the parameters. These are the logarithmic, inverse, and square root transformations. In the analysis of $\hat{a}$ versus a data from the ECIS, the logarithmic transformation of $\hat{a}$ has always been necessary. The logarithmic transformation of crack size has also been needed to obtain linearity. Further, the inverse transformation, $1/a$, is also capable of providing linearity for the characteristic curvature often seen in the ECIS $\ln \hat{a}$ versus $\ln a$ plots. The

appropriateness of particular transformations for different responses and patterns of changing scatter are discussed in Wasserman and Kutner [17].

> Note: Given linearity and equal variance, the assumption of normality of the residuals is usually acceptable. Outliers may cause rejection in a statistical test of normality in which case the quality and effect of the outliers must be judged individually. Distributions other than normal can be implemented, but a family of distribution must be specified to implement the maximum likelihood analysis. Since normality has been generally acceptable in NDE data for which the distribution assumption could be tested, only normally distributed residuals are considered in this report.

The guidelines for applying transformation are as follows:

1. Transform the variable with the greatest range (maximum/minimum), since transformations have little effect over a narrow range.
2. Determine if linearity and homoschedasticity are acceptable.
3. Transform the other variable, if necessary.
4. Determine if linearity and homoschedasticity are acceptable.
5. If transformations do not provide an acceptable degree of linearity and homoschedasticity, segment the data and repeat the steps for the relevant segments.

The application of this process to early eddy current $\hat{a}$ versus a data indicated the necessity to transform the inspection response to stabilize the variance and the crack size to produce a broader range of linearity. The general versatility of the log-log model led to the decision to program the POD(a) analysis in these terms. Note, however, there are sets of evaluation data for which the logarithmic transformations are not necessary.

In the ECIS evaluations, the variability in response increases with crack size due to the multiplicative nature of calibration. Since the logarithmic transformation tends to stabilize the standard deviation, the POD(a) modeling for the ECIS has always been in terms of ln $\hat{a}$. The standard deviation of the residuals of the transformed responses (ln $\hat{a}$) has not always been constant across the range of crack size in the demonstration specimens. The nonhomogeneity of variance was usually attributable to increased scatter at the smallest crack sizes but, on occasion, there was increased scatter when crack sizes reached the probe size.

The increased scatter at the small sizes was discussed in subsection 2.2.2. If the target a_{90} value is in the range of crack sizes with the larger standard deviation, it may be necessary to partition the data before analysis. If the target a_{90} values are outside the range of the increased standard deviation of ln $\hat{a}$ residuals, the change in scatter can be subjectively evaluated. As noted previously, the composite residual standard deviation, σ_{δ} will be larger (conservative).

When the size of the inspected crack is large compared to the size of the probe, the average signal response tends to approach or reach a maximum, and the scatter in the ln $\hat{a}$

responses may increase. Figure 8 is an example of such data from ECIS inspections of titanium bolt holes. Because $\hat{a}$ is not continuing to increase with a, there is an upper limit to $\hat{a}_{dec}$. Crack sizes beyond some upper limit determine the range of applicability only of the specific inspection. Increased scatter in $\hat{a}$ data from cracks above the upper limit should not be included in the analysis to determine $\hat{a}$ thresholds for crack sizes below the limit. If such large cracks are the cause of rejecting homogeneity of variance, they should be excluded from the POD(a) analysis.

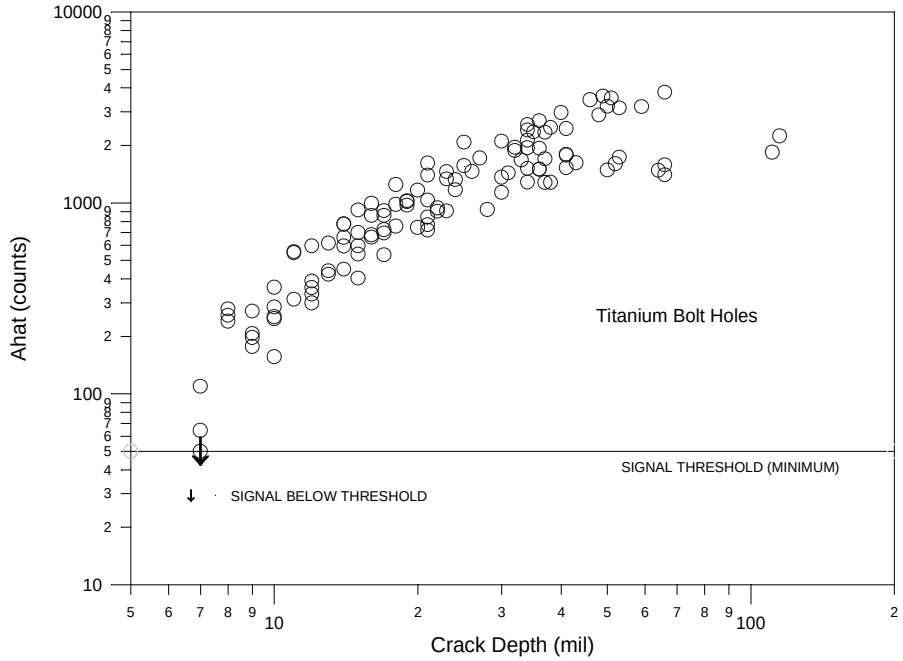


Figure 8. Example Non-linear ln $\hat{a}$ versus ln a Response for Titanium Bolt Holes

In Figure 8, $\hat{a}$ thresholds could be determined for target a_{90} values up to about 30 mils. Because the average ln $\hat{a}$ is not increasing for $a > 30$, $\hat{a}$ thresholds for cracks greater than 30 would be the same as the threshold for $a = 30$. The increase in scatter in data from cracks greater than 30 would be used only to bound the applicable range of the results and should not influence the detection thresholds for smaller crack sizes. In the data of Figure 8, the detection threshold for $a_{90} = 40$ mil might allow cracks of about 60 mil to be undetected.

The logarithmic transformation of $\hat{a}$ was always necessary in the analysis of the ECIS data to equalize the variance, but the logarithmic transformation of crack size was also necessary to linearize the relation. Logarithmic transformations of both $\hat{a}$ and a often produced data for which the analysis assumptions were satisfied. When the log $\hat{a}$ versus log a relation has not been linear, the departures from linearity exhibited a rather consistent pattern. Figure 9 is representative of such non-linear $\hat{a}$ versus a behavior. The log $\hat{a}$ responses are concave downward with increasing a and may approach a non-saturation maximum in the mean, an example of which is shown in Figure 8. (On occasion, an S-shaped response due to a further flattening of response at small crack sizes

was observed. As noted earlier, there is no assurance that such responses are not due to the difficulty in characterizing the size of smaller cracks.)

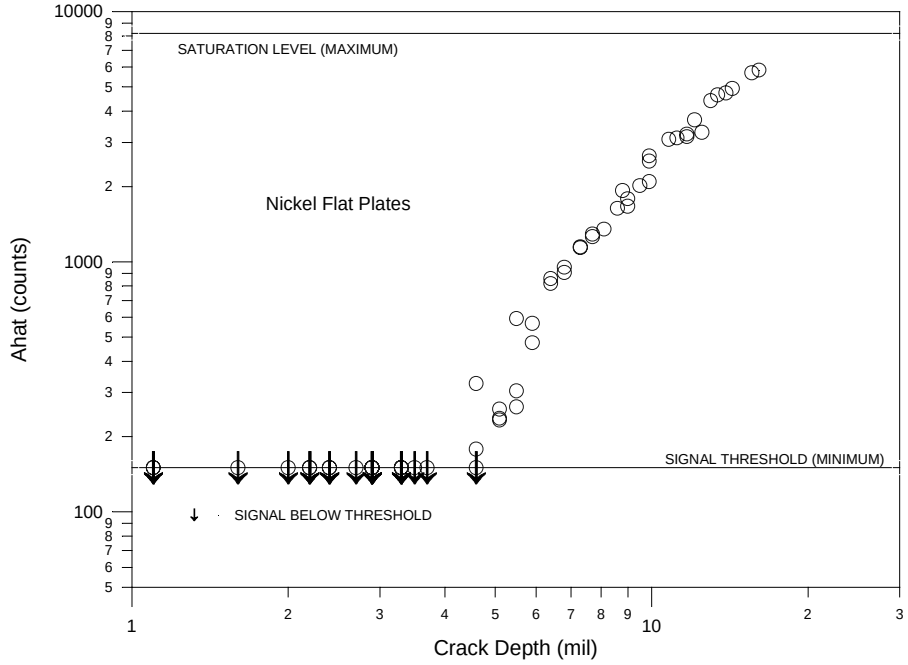


Figure 9. Example Non-linear ln $\hat{a}$ versus ln a Response for Nickel Flat Plates

The logarithmic transformation of crack size is one of the recommended transformations to account for curvature. In the data of Figure 9, the logarithmic transformation was not adequate. A stronger transformation is the inverse crack size. Figure 10 presents ln $\hat{a}$ versus 1/a for the data of Figure 9. To avoid compression in the crack size range of interest, the below minimum $\hat{a}$ responses for cracks smaller than 2.5 mil are not included in Figure 10. The straight line fit shown on the plot was calculated using the standard $\hat{a}$ versus a analysis. The linearity, equal variance, and normality assumptions were not rejected. Figure 11 presents the fit obtained from inverse transformation on the ln $\hat{a}$ versus ln a data of Figure 9.

When the data of this example were originally analyzed, it was necessary to partition the data into small and large crack regions to obtain reasonable linearity in the ln $\hat{a}$ versus ln a relation. Figure 12 compares the a_{90} versus decision thresholds from the original POD(a) analysis of these data using the partitioned data sets and that obtained from the analyzing ln $\hat{a}$ versus 1/a across the entire range of crack sizes. In this example, the agreement is excellent.

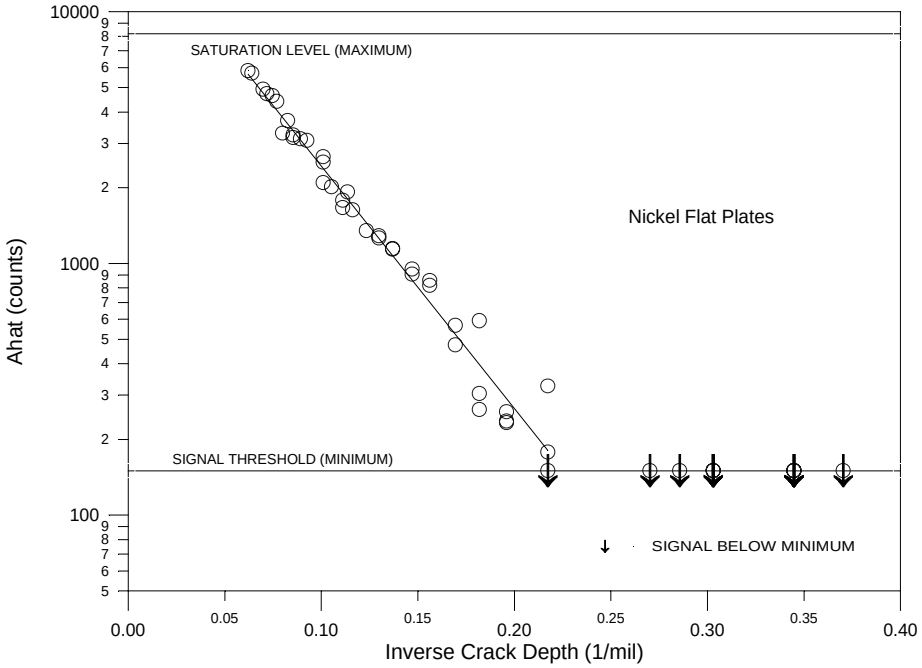


Figure 10. Linear Fit to ln $\hat{a}$ versus 1/a for Nickel Flat Plates

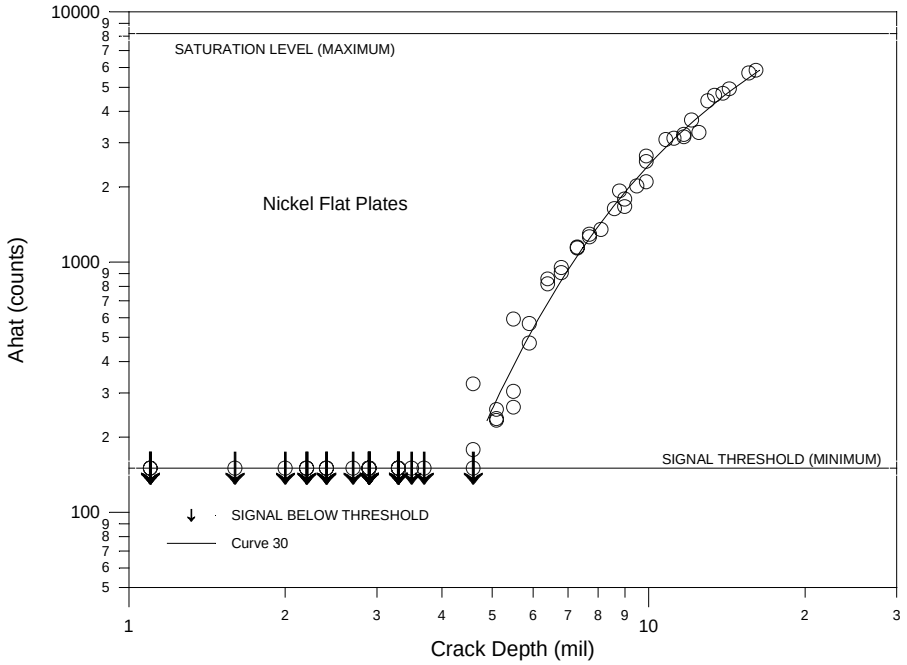


Figure 11. Fit from 1/*a* Transformation on ln $\hat{a}$ versus ln *a* for Nickel Flat Plates

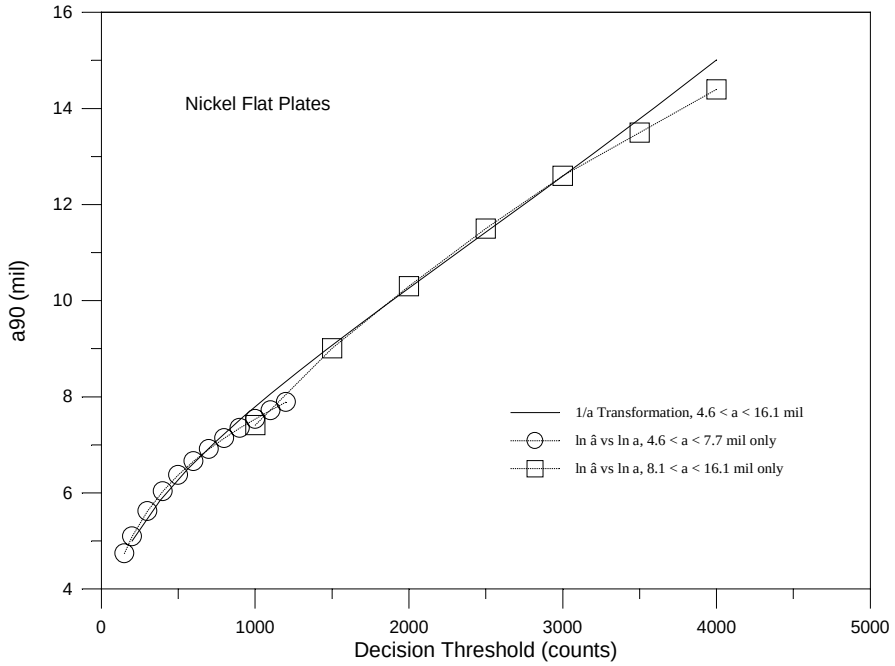


Figure 12. a_{90} versus Decision Threshold Comparing Results from ln $\hat{a}$ versus 1/*a* to Original ln $\hat{a}$ versus ln *a* Analysis

The ln $\hat{a}$ versus ln a titanium bolt hole data of Figure 9 are not linear. To obtain linearity in the target a_{90} range of interest, the original analysis was restricted to crack sizes less than 20 mil. When the analysis is applied in terms of ln $\hat{a}$ versus 1/a, the relation is acceptably linear for crack sizes less than 50 mil. As noted earlier, the POD(a) analysis for this data set should be limited to about 30 mil because of the change in response for large cracks. Figure 13 shows the linear fit obtained from the POD(a) program for the ln $\hat{a}$ versus 1/a data. Figure 14 shows the fit on the ln $\hat{a}$ versus ln a plot of Figure 9. Figure 15 compares the a_{90} versus decision threshold plots from the original analysis and that obtained from the inverse transformation.

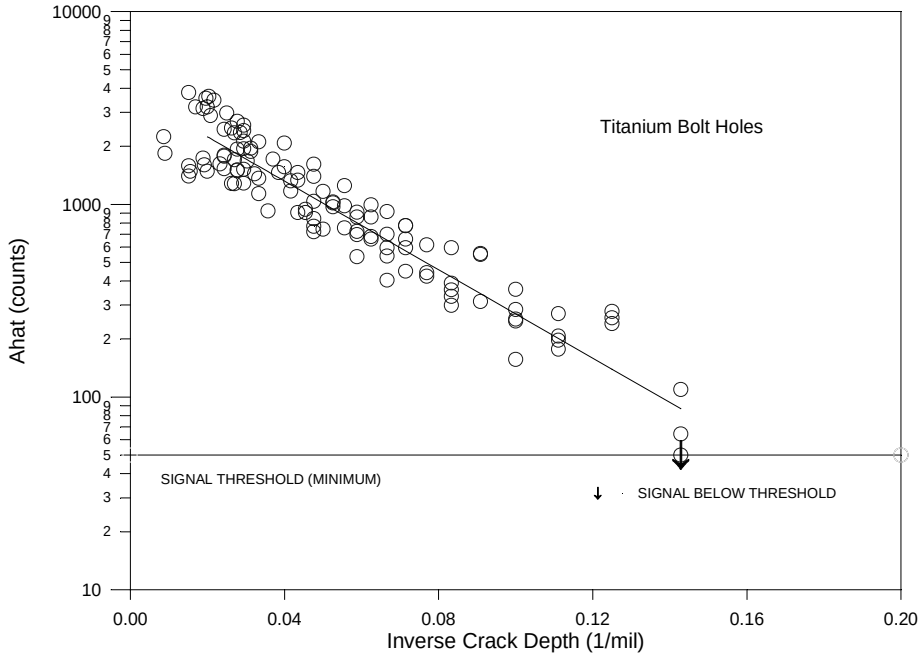


Figure 13. Linear Fit to ln $\hat{a}$ versus 1/a for Titanium Bolt Holes

The inverse transformation has been successfully applied to several data sets that display the characteristic curvature in the ln $\hat{a}$ versus ln a plot. The crack size range of applicability for the ln $\hat{a}$ versus 1/a analysis was greater than that of the ln $\hat{a}$ versus ln a analysis, but the a_{90} results at the largest crack sizes must be carefully evaluated. Because the 1/a transformation can greatly compress the analysis scale for the large crack sizes, the inverse transformation can generate unrealistically high a_{90} values. This happens only at the extreme upper end of the crack size range when the total range of crack sizes is large. In this situation, the unrealistic a_{90} values are readily detectable. Under no circumstance should the results be extrapolated beyond the largest crack size in the specimen set.

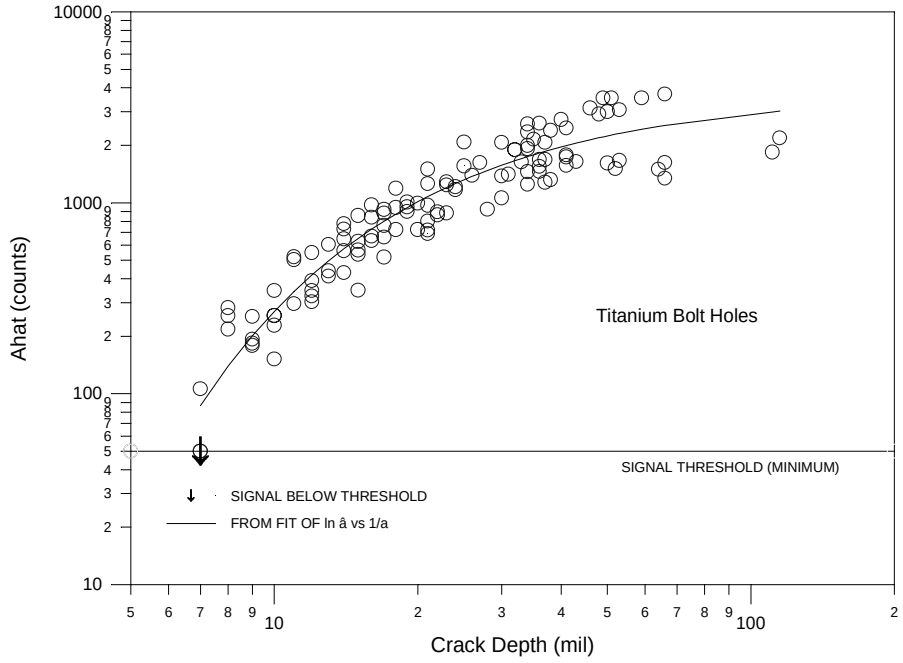


Figure 14. Fit from 1/*a* Transformation on ln $\hat{a}$ versus ln *a* for Nickel Flat Plates

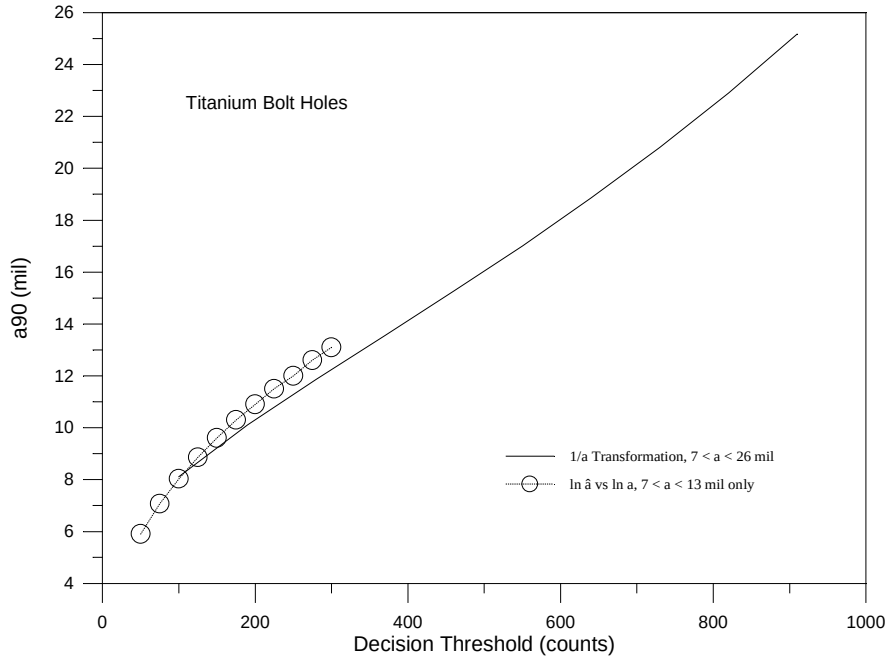


Figure 15. a_{90} versus Decision Threshold Comparing Results from ln $\hat{a}$ versus 1/*a* to Original ln $\hat{a}$ versus ln *a* Analysis

The problems with nonlinearity and increased scatter at the small crack sizes are not resolved by transformations. A spline fit approach may handle an arbitrary nonlinear $\hat{a}$ versus a relation but would not account for changes in the standard deviation of residuals that have been observed at the small crack sizes. As noted in subsection 2.1.3, at least some, if not all, of the erratic behavior in the $\hat{a}$ versus a response at small crack sizes is attributable to mis-sized cracks. It is doubtful that a_{90} values would be achievable in the small crack range when the small cracks introduce an S-shape to the inspection response and a significant increase in the scatter about the average response. Further, the nonrepresentative small crack responses should not be included in an analysis for larger target a_{90} values. Therefore, excluding the erratic responses from the small cracks should have no practical effect on the characterization of POD(a) at larger crack sizes. If a_{90} values are needed for very small cracks, methods for generating representative cracks and characterizing their sizes must be developed.

It might be noted that consideration was given to introducing a quadratic fit for the ln $\hat{a}$ versus ln a data when the data fail the linearity test. The POD(a) function from a quadratic fit would not have the form of a cumulative normal distribution function, and the results would require tabular formats. Quadratic fits were made for a number of data sets that exhibited the characteristic downward concave shape with increasing crack size. Although some of the fits were good, in general the quadratic fit was not judged better than the fit obtained from the inverse crack size transformation. Eliminating the small cracks and partitioning into ranges would still be necessary, and added constraints would be needed to prevent occasional nonconservative interpretations of the analysis. For example, in one of the test data sets, the response was linear over much of the range of crack sizes, and the resultant quadratic fit was concave upward. The mean response had a midrange conservative bias but a nonconservative bias in the larger crack size region.

2.3 Pass/Fail Analysis

When the recorded response of an NDE system is a yes or no decision only as to the presence or absence of a flaw, the parameters of the POD(a) function must be estimated directly from the data. This is accomplished by assuming a model for the POD(a) function and determining the values of the parameters that maximize the likelihood (probability) of obtaining the finds and misses that were obtained in the inspections of the demonstration. The asymptotic properties of the maximum likelihood estimates of the parameters of the POD(a) functions are known and are used to place confidence bounds on the estimates of POD(a). The mathematics of the procedure are fully explained in Mil-HDBK-1823 [4] and Berens [5] but a modification will be introduced here to reflect the continued use of the $a_{90/95}$ value as the dominant objective of a POD analysis.

2.3.1 Pass/Fail Model Formulation

Let a_i represent the size of the i^{th} flaw and Z_i represent the result of the inspection: $Z_i = 1$ if the flaw was found (pass) and $Z_i = 0$ if the flaw was missed (fail). Assume that POD(a_i) is the equation for the probability of detecting a flaw of size a_i during the inspection. The

likelihood of obtaining a specific set of (a_i, Z_i) results when inspecting the specimens is as follows:

$$L(\boldsymbol{\theta}) = \Pi\, [POD(a_i)]^{Z_i}\, [1- POD(a_i)]^{1-Z_i}\,, \qquad (22)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of the parameters of the POD(a) function. Values of $\theta_1, \theta_2, \ldots, \theta_k$ are determined to maximize $L(\boldsymbol{\theta})$. For typical POD(a) models, it is more convenient to perform the analyses in terms of logarithms as follows:

$$\ln L(\boldsymbol{\theta}) = \Sigma\, Z_i \ln POD(a_i) + \Sigma\, (1-Z_i) \ln [(1- POD(a_i)] \qquad (23)$$

The maximum likelihood estimates are given by the solution of the k simultaneous equations:

$$\delta \ln L(\boldsymbol{\theta}) / \delta\ \theta_i = 0,\ \ i = 1, \ldots, k \qquad (24)$$

In general, an iterative solution will be required to solve equations 24.

Any monotone increasing function between zero and one can be used for POD(a). However, an early study of data with multiple inspections per crack [13] indicated that the log odds model was more generally applicable than the others investigated. Further, the assumptions leading to a cumulative lognormal model for the POD(a) function for $\hat{a}$ versus a data have often been verified for eddy current data. The log odds and cumulative lognormal models are equivalent in a practical sense in that the maximum difference in POD(a) between the two for fixed location and scale parameters is about 0.02, which is well within the scatter from repeated determinations of a POD(a) capability. To be consistent with the $\hat{a}$ versus a analysis, the computer program of Mil-HDBK-1823 [4] is based on a cumulative lognormal equation. However, on occasion the maximum likelihood equations based on the cumulative lognormal equation could not be solved when a solution using the log odds equations was possible. Accordingly, both equations were programmed in the updated POD program. Other models for the POD(a) function may be appropriate but, if preferred, would require a different computer implementation.

Repeating equation 4, the cumulative lognormal equation for the POD(a) functions is given by:

$$POD(a) = \Phi[(\ln a - \mu)/\sigma], \qquad (25)$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function. The log odds model for the POD(a) function is as follows:

$$POD(a) = \{1 + \exp - [(\pi/\sqrt{3})(\ln a - \mu)/\sigma]\}^{-1} \qquad (26)$$

Either equation 25 or 26 is substituted in Equations 24. The $\hat{\mu}$ and $\hat{\sigma}$ are determined so as to maximize $L(\mu,\sigma)$, the likelihood of obtaining the observed inspection results. Note that $POD(\mu) = 0.5$ for both models. The σ is a scale parameter that determines the degree of steepness of the POD(a) function. A negative value of σ is not contradictory but, for a negative σ, the POD(a) function will decrease with increasing a.

As noted, there are occasions when Equations 24 do not converge. No solution will be obtained if the sizes of found cracks do not overlap with the sizes of missed cracks. Little information is obtained from cracks that are so large they are always found or so small they are always missed. More overlap is needed for the cumulative lognormal model than for the log odds model. It is also possible to obtain negative estimates of σ from erratic data sets. Results of this nature are due to the wrong range of crack sizes in the demonstration or to an inspection process that is not under proper control. When the crack sizes in the specimens are not in the range of increase of the POD(a) function, the effective sample size is smaller and the effect is reflected in larger standard deviations of the sampling distributions of the parameter estimates and, thus, wider confidence bounds.

2.3.2 Confidence Bounds on a_{90}

Damage tolerance analyses are driven by the single crack size characterization of inspection capability for which there is a high probability of detection. Typically, the one number characterization of the capability of the NDE system is expressed in terms of the crack length for which there is 90 percent probability of detection, a_{90}. But a_{90} can be estimated only from a demonstration experiment and there is sampling uncertainty in the estimate. To cover this variability, an upper confidence bound can be placed on the best estimate of a_{90}. The use of an upper 95 percent confidence bound has become a de facto standard for this characterization of NDE capability, which is intended to be conservative from the viewpoint of damage tolerance analyses.

The estimated crack size for which there is 95 percent confidence that at least 90 percent of cracks will be detected by the system is known as the 90/95 crack size, $a_{90/95}$. In a similar fashion, it has become customary to refer to the best estimate of a_{90} as $a_{90/50}$, or the 90/50 crack size. As previously noted, in the $\hat{a}$ versus a analysis the $a_{90/50}$ value is used to characterize capability, since these estimates tend to be stable for data of this nature. However, the $a_{90/50}$ values have much more sampling variability in the pass/fail analysis, and the confidence bound is needed to account for this scatter.

In the pass/fail analysis described in Mil-HDBK-1823 [4] and Berens [5] and implemented in the recommend analysis computer program of Mil-HDBK-1823 [4], the confidence limit for a_{90} was calculated from the confidence bound on the entire POD(a) function. In this approach, there is 95 percent confidence that the entire POD(a) function lies above the calculated bound. The $a_{90/95}$ value was determined as the crack size at which this calculated bound crosses 0.90. Calculating $a_{90/95}$ using this approach introduces conservatism in an estimate of any single POD(a) value because of the insistence that the entire POD(a) function must lie above the bound. An alternate approach is to select a

single value of POD(a), such as 0.90, and place an upper confidence bound on a only at the single value. This procedure is known as a point-by-point confidence bound and produces shorter confidence intervals at any single value of POD(a).

The point-by-point confidence bound was deliberately not implemented in Mil-HDBK-1823 [4]. At that time, the use of the entire POD(a) function was being promoted for use as the characterization of NDE capability, so the decision was made to implement the confidence bound for the entire POD(a) function. However, the use of $a_{90/95}$ has not diminished and, today, is the most commonly used characterization of NDE capability. For this reason, the less conservative point-by-point confidence bounds on POD(a) will be programmed in the update of the POD software. These are valid confidence bounds for any one POD value but not for the entire POD(a) curve.

2.3.2.1 Confidence Bounds for the Cumulative Lognormal Model

Assume that the POD(a) function is being modeled by the cumulative lognormal distribution function, equation 25. Let a_p represent the crack size for which POD(a_p) = p. Then, if μ and σ are known exactly, the following applies:

$$a_p = \exp(\mu + z_p\, \sigma), \tag{27}$$

where z_p is the p^{th} percentile of the standard normal distribution, and $p = \Phi(z_p)$. For example, $\Phi(1.282) = 0.90$ so that $a_{90} = \exp(\mu + 1.282\ \sigma)$. But μ and σ are never known exactly. Rather, they are estimated from the pass/fail data of an NDE demonstration. The best estimate of the p percent detectable crack size is as follows:

$$a_p = \exp(\hat{\mu} + z_p\, \hat{\sigma}), \tag{28}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimates of μ and σ, respectively. Because $\hat{\mu}$ and $\hat{\sigma}$ are maximum likelihood estimates, they are asymptotically normally distributed. Thus, the percentile $X_p = \hat{\mu} + z_p\hat{\sigma}$ has an asymptotic normal distribution with mean and standard deviation given by the following:

$$M(X_p) = \mu + z_p\, \sigma \tag{29}$$

and

$$SD(X_p) = [V_{11} + 2\, z_p\, V_{12} + z_p^{\,2}\, V_{22}]^{0.5}, \tag{30}$$

where V_{11} is the variance of $\hat{\mu}$, V_{22} is the variance of $\hat{\sigma}$, and V_{12} is the covariance of V_{11} and V_{22}. The variances and covariance of $\hat{\mu}$ and $\hat{\sigma}$ are calculated and contained in the routine output of the POD(a) maximum likelihood estimation program. An upper 100 q percent confidence bound for the true $\mu + z_p\, \sigma$ is given by $X_p + z_q\, SD(X_p)$, where z_q is the q^{th} percentile of a standard normal distribution. Let the notation $a_{p/q}$ denote a confidence

level of q that at least p percent of the cracks of size $a_{p/q}$ will be detected. Then $a_{p/q}$, is estimated by the equation:

$$a_{p/q} = \exp\left[\hat{\mu} + z_p \hat{\sigma} + z_q \, SD(X_p)\right] \tag{31}$$

For the particular case of the 90/95 reliably detected crack size, the following applies:

$$a_{90/95} = \exp[\hat{\mu} + 1.282 \hat{\sigma} + 1.645 \, SD(X_p)] \tag{32}$$

Since $z_{0.5} = 0$ for q = 0.5 (50 percent confidence), the best estimate of a_p in equation 28 is the $a_{p/50}$ estimate. In particular, $a_{90/50} = \exp[\hat{\mu} + 1.282 \hat{\sigma}]$.

As an example, Figure 16 presents a lognormal POD(*a*) fit with confidence bounds to pass/fail data from a specimen set of nickel flat plates. The $a_{90/95}$ value for the point by point confidence bound is 83 mil. The $a_{90/95}$ value calculated from the bound on the entire POD(*a*) function is 243 mil. Both are valid confidence bounds, but they must be correctly interpreted. It is not valid to use the confidence limit for more than one POD value from the 95 percent bound on each POD value. However, there is 95 percent confidence that at least p percent of all cracks greater than $a_{p/95}$ will be detected for all values of p when $a_{p/95}$ is obtained from confidence bound on the entire POD(*a*) function. Since the cracks that were greater than 40 mil were always detected the tighter bound at POD(*a*) = 0.9 appears more reasonable.

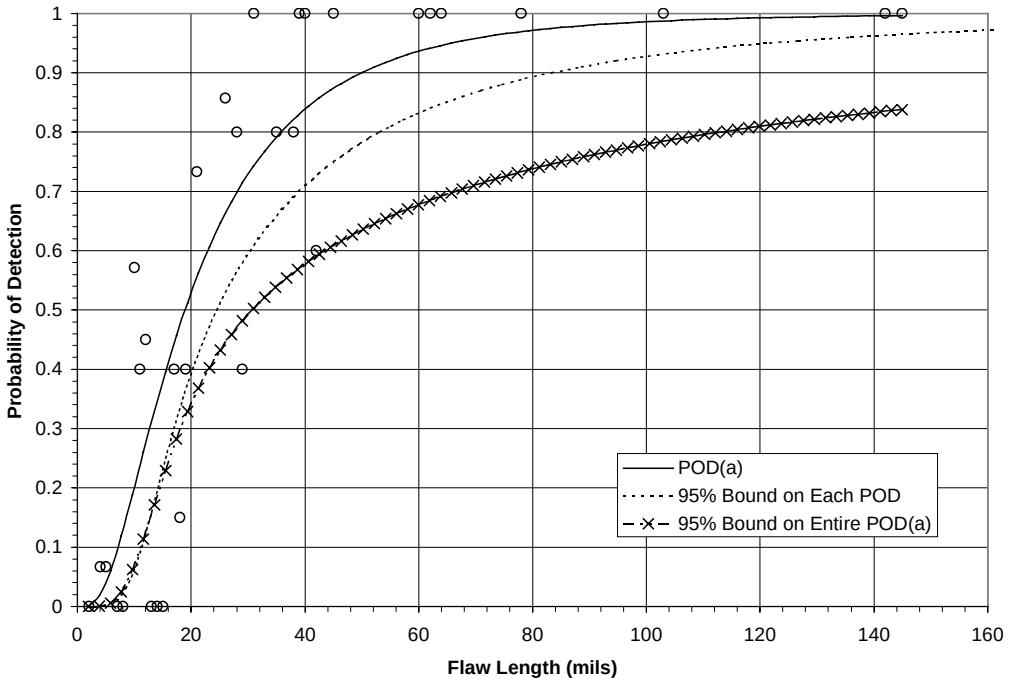


Figure 16. Comparison of POD(*a*) Confidence Bounds from Pass/Fail Analysis

2.3.2.2 Confidence Bounds for the Log Odds Model

Assume that the functional form of the POD(a) equation is the log odds model given by the following:

$$\mathrm{POD}(a) = \{1 + \exp[\, -\pi\, (\ln a - \mu) / \sigma\sqrt{3}]\}^{-1}. \qquad (33)$$

If a_p represents the crack size for which POD(a_p) = p and if μ and σ are known exactly, then

$$a_p = \exp(\mu + C_p\, \sigma) \qquad (34)$$

where $C_p = -\sqrt{3}\ \ln[(1\text{-}p)/p] / \pi$. The point-by-point confidence bounds for a_p values are calculated analogous to the development of equations 28 through 31. The best estimate of the p percent detectable crack size is given by:

$$a_p = \exp(\hat{\mu} + C_p\, \hat{\sigma}), \qquad (35)$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimates of μ and σ, respectively. Again, $\hat{\mu}$ and $\hat{\sigma}$ are asymptotically normally distributed. Thus, $X_p = \hat{\mu} + C_p\hat{\sigma}$ has an asymptotic normal distribution with mean and standard deviation given by the following:

$$\mathrm{M}(X_p) = \mu + C_p\, \sigma \qquad (36)$$

and

$$\mathrm{SD}(X_p) = [V_{11} + 2\, C_p\, V_{12} + C_p^{\,2}\, V_{22}]^{0.5}, \qquad (37)$$

where V_{11} is the variance of $\hat{\mu}$, V_{22} is the variance of $\hat{\sigma}$, and V_{12} is the covariance of $\hat{\mu}$ and $\hat{\sigma}$. Estimates of the variances and covariance of $\hat{\mu}$ and $\hat{\sigma}$ are calculated and contained in the routine output of the POD(a) maximum likelihood estimation program. An upper 100 q percent confidence bound for the true $\mu + C_p\, \sigma$ is given by $X_p + z_q\, \mathrm{SD}(X_p)$ where z_q is the q^{th} percentile of a standard normal distribution. Using the notation $a_{p/q}$ to denote a confidence level of q that at least p percent of the cracks of size $a_{p/q}$ will be detected, the following equation results:

$$a_{p/q} = \exp\,[\hat{\mu} + C_p\, \hat{\sigma} + z_q\, \mathrm{SD}(X_p)]. \qquad (38)$$

The 90/95 reliably detected crack size is given by the following:

$$a_{90/95} = \exp[\hat{\mu} + 1.211\, \hat{\sigma} + 1.645\, \mathrm{SD}(X_p)]. \qquad (39)$$

Since $z_{0.5} = 0$ for q = 0.5 (50 percent confidence), the best estimate of a_p in equation 35 is the $a_{p/50}$ estimate. In particular, $a_{90/50} = \exp[\hat{\mu} + 1.211\, \hat{\sigma}]$.

2.3.3 Goodness of Fit Test of Pass/Fail Models

In the $\hat{a}$ versus a analysis, the validity of the lognormal cumulative distribution function as a model for POD(a) can be verified by tests using the recorded $\hat{a}$ values. In the pass/fail analysis, the only inspection response is the presence or absence of a flaw, and any tests of model validity must be based on the predicted and observed numbers of finds. Statistical goodness of fit tests have been devised for data of this categorical nature [18]. These tests compare the expected number of finds in ranges of crack size from the POD(a) fit with the observed number of finds in the data using the chi square distribution. The statistical validity of the tests depends on having at least a few cracks in each of the intervals.

The Hosmer-Lemeshow test [18] is representative of such goodness of fit tests for categorical data and was applied to several relatively large pass/fail data sets. Each of the specimen sets contained at least 100 cracks, but none of the data sets had a sufficient number of cracks in the range of increase of the POD(a) function to make the assumptions of the test reasonably valid. In general, specimen sets tend to have a disproportionately large number of cracks for which detection is very likely, for example, those for which POD is greater than 0.9. For these very common data sets, the assumption required by the tests will not be reasonable.

Since there is no assurance that the assumptions required by a goodness of fit test for pass/fail data would be reasonably acceptable or evaluated, the decision was made not to include a goodness of fit test in the updated POD computer program. Goodness of fit of the cumulative lognormal distribution function or the logs odds model for the POD(a) equation must be judged subjectively by comparing the fit to the observed data. If a different model is preferred, the POD computer program of this report cannot be used.

Section 3
POD Program Update

There are two POD computer programs referenced in Mil-HDBK-1823 [4] for analyzing the two types of response data from NDE capability demonstrations. These programs, AHAT and PF, are written using the computer and software capabilities that were available in the early 1980's. The analyses are coded in FORTRAN with a fixed field (card image) type input from a text file. The user interface with the programs is DOS-based and not user friendly by today's standards. The output from the analyses is written to three text files:

- A summary of the analysis comprising the estimates of the parameters of the model, the variance-covariance matrix of the parameter estimates, and the a_{50}, a_{90}, and $a_{90/95}$ values for selected thresholds.
- The estimated POD(a) function and 95 percent confidence bound for one decision threshold.
- The calculated and observed POD at the crack sizes in the demonstration specimen set.

All plots summarizing or presenting the results were generated from these text files using external plotting routines. Evaluations of the model were conducted independently using the input and output text files.

A prime objective of the study of this report was to update the POD computer programs to take advantage of current computer capabilities. To make the updated program usable on a broad range of computer platforms, the ubiquitous Microsoft Excel® was selected as the primary user interface. The analysis codes were rewritten in C^{++}, statistical tests of the $\hat{a}$ versus a assumptions were made an integral part of the analysis, and selected plots were added as standard options. This section of the report describes the use of the Excel interface, the additions that were made to the analyses, and the standard output additions that were built into the program. Volume 2 of this report is a standalone users manual for the updated POD programs and presents a detailed use of the program.

3.1 User Interface

A workbook of spreadsheets is a natural interface for the POD calculation programs because of the wealth of additional workbook features that are available to change and supplement the analyses. Ease of generating and modifying input files for different analyses, the designation of selected pages for standard output, the availability of plotting, and access to new worksheets for additional comparisons or analyses of the basic data set all support the use of the workbook format. In particular, Microsoft Excel 97 was selected as the interface between the user and the updated POD analysis programs because all Air Force users and most others will have access to Excel.

For the POD analyses, the workbook is controlled though a window that opens and closes workbooks, calls for recalculations using the analysis and data as indicated in the

workbook, and specifies which of the standard charts to generate. This window controls the program entitled POD which is the interface between the C^{++} programs that perform the maximum likelihood calculations and the data and output of the workbook. The POD program is the controlling program and is external to the Excel workbook. This subsection focuses on the input to an analysis and the interface between POD and the Excel workbook.

The input for a POD analysis is contained on two sheets of the Excel workbook: the **Data** sheet and the **Info** sheet. The **Data** sheet contains the flaw sizes and results from the capability demonstration inspections. The **Info** sheet contains identification, modeling, and data information that will used by the POD program. Although a workbook can be repeatedly opened, the following discussion assumes that a data set is being prepared for its first use by POD.

The basic data input to POD is an Excel spreadsheet. Row one of the spreadsheet is a header row containing the names of the columns. Each succeeding row contains the identifying information, crack size, and all inspection results for a single crack of the specimen set. The POD program uses only the columns that contain the crack size, as designated by the column name, and the inspection results. The POD program will request the column number that contains the first inspection result, and all of the inspection results must be contiguous.

As an example, Figure 17 presents part of a **Data** spreadsheet that contains the first 20 cracks of a set for an $\hat{a}$ versus a analysis. The first three columns of this example **Data** sheet contain identifying information for the cracks. Both crack length and depth are listed for the cracks, and the analyst will specify which is to be used in the analysis. Each crack in the example data sheet has been inspected twice and the inspection designations are I11 and I21. Note that some of the inspection sites do not contain cracks. These can be removed later.

When a new Excel workbook containing the inspection results is opened by POD, the program first ensures that the data sheet is named **Data**. An **Info** sheet is then initiated and the **Info** sheet contains the minimum data needed to perform the requested POD analysis. Whenever POD opens a workbook, the POD model of the **Info** sheet is recalculated to ensure that all results and plots were generated from the current setup. (Note that a workbook can be repeatedly opened by POD. The recalculation is necessary to ensure that the output in a workbook was calculated from the information on the **Info** and **Data** sheets.) After a workbook has been opened and recalculated, the data to be analyzed can be modified and the **Info** sheet can be changed to provide a more complete description of an analysis run and to change analysis parameters or models. Such changes are made on the **Info** sheet, a sample of which is presented in Figure 18 for an ahat analysis.

S/N	Surface	Theta	Length	Depth	I11	I21
1	top	0	35	17	322	383
2	blank	blank	blank	blank	0	0
3	top	75	28	13	284	300
4	blank	blank	blank	blank	0	0
5.1	top	135	29	15	224	285
5.2	top	315	28	13	272	276
6	top	315	34	16	307	360
7	blank	blank	blank	blank	0	0
8	bottom	240	6	3	0	0
9	top	90	27	13	322	327
10	bottom	315	26	12	223	254
11	top	0	20	9	217	291
12	top	0	33	16	378	548
13	blank	blank	blank	blank	0	0
14	bottom	315	26	12	271	254
15	bottom	80	32	15	417	508

Figure 17. Partial **Data** Worksheet for an $\hat{a}$ versus a Analysis Using POD

ID INFO			
Title:	Titanium Small Bolt Holes		
Subtitle:	Example line 2		
Subtitle:	Example line 3		
Subtitle:	Example line 4		
Subtitle:	Example line 5		
Subtitle:	Example line 6		
FLAW INFO			
Flaw Name:	Depth		
Flaw Units:	mil		
Flaw Transform:	log		
SIGNAL INFO			
Insp Start:	F		
Insp Units:	counts		
Insp Transform:	log		
Signal Min:	50		
Signal Max:	1000		
ANALYSIS			
Analysis:	Ahat		
Version:	POD 3.0		
Thresholds:	50	500	
POD Threshold:	150		
POD level:	90		
Confidence:	95		

Figure 18. Example **Info** Sheet for an $\hat{a}$ versus a Analysis Using POD

The Title line and the first five Subtitle lines of the **Info** sheet are printed on all output of an analysis. As many subtitle lines as desired can be inserted in the **Info** sheet and all will be printed on the output sheet entitled **Results**. For an $\hat{a}$ versus a analysis, the **Info** sheet defines:

- the units and transformation of the crack sizes in the analysis
- the units and censoring values for $\hat{a}$
- the analysis (ahat in this example) to be performed by POD
- the range of $\hat{a}$ decision thresholds for the calculation of the a_{90} versus $\hat{a}_{dec}$ plot
- the $\hat{a}$ decision threshold for a plot of the POD(a) function with confidence bound.

(The same **Info** sheet is used for the pass/fail analysis, but only the parameters relevant to the pass/fail analysis are used.)

The data to be used in an analysis are selected on the **Data** worksheet. Note first that POD provides an easy sort by crack size of the data in the worksheet. See POD window, **Tasks**, **Sort by size**. But note also that the **Data** worksheet is an Excel spreadsheet and any spreadsheet functions can be performed. It is necessary to maintain the inspection results in contiguous columns starting in the column as indicated on the **Info** worksheet. POD will use all cracks that have a clear (uncolored) background. To eliminate cracks from an analysis, select them and add a highlight background color. This selection process makes it very convenient to select particular inspections of all cracks or to exclude ranges of crack size by adding any of the background colors to the unwanted crack responses.

The first time POD opens a **Data** worksheet, the program estimates the parameters and creates a **Results** and a **Residuals** worksheet of output based on the initiating analysis conditions. This analysis will seldom be the final analysis. To obtain plots, analysts open the chart window of POD where five standard plots are available. The data for three of these plots are in the **Residuals** worksheet and POD will create two additional worksheets for the data of the other two plots. Plots must be requested the first time, but subsequent recalculations of the POD analysis will regenerate the previously requested plots. Descriptions of the output worksheets and plots are presented in subsection 3.3.

The POD program will name, use, and control a maximum of 11 sheets of the workbook. Non-POD columns of the **Info** and **Data** worksheets can be used at the discretion of the analyst, but the other named worksheets of POD will be recreated for new analysis runs. When new worksheets are introduced, they must have names that are different from those assigned by POD.

3.2 Analysis Additions

The original AHAT computer program calculated the parameters for a linear fit to ln $\hat{a}$ versus ln a data, produced a table of a_{50}, a_{90}, and $a_{90/95}$ values for defined decision thresholds, and generated a text file for POD(a) and 95 percent confidence bound for one decision threshold. The original PF program calculated the parameters and a_{50}, a_{90}, and $a_{90/95}$ values for the lognormal cumulative distribution model for POD(a). It also

generated a text file for the estimate of POD(*a*) and its 95 percent confidence bound. The updated POD program performs these and additional analyses. In particular, provisions were added to perform the analyses on transformed data and to test the basic assumptions of the $\hat{a}$ versus *a* analysis.

3.2.1 Transformations

The POD(*a*) calculations of the original AHAT computer program were performed in terms of the logarithms of crack size and inspection response. Similarly, the POD(*a*) calculations of the original PF computer program were performed in terms of the logarithms of crack size. These particular transformations of the basic input data were selected because they were found to provide a generally acceptable model for the POD(*a*) function. As discussed in subsection 2.2.3, other transformation of the basic input data might be preferred in selected applications. Accordingly, the updated POD program provides the capability to perform the analyses using transformations other than the logarithmic.

As shown in the example **Info** sheet of Figure 18, the Flaw Transform and Insp Transform rows are used to specify the transformations, if any, that will be used in the analysis. The Insp Transform line is applicable only for an $\hat{a}$ versus *a* analysis. The transform is defined by the entry in column B for the transform line and the possible entries are as follows:

- No entry will result in the default analysis that is based on the natural logarithm transform. Such analyses are identical to those of the existing POD programs.
- An entry of *log* will result in an analysis based on the natural logarithm transformation. Such analyses are identical to those of the existing POD programs. No entry and "log" will yield identical results.
- An entry of *none* will result in the data being analyzed exactly as they are recorded on the **Data** sheet – i.e., without any transformations.
- An entry of *inverse* will perform the analysis in terms of 1/a or 1/$\hat{a}$.
- An entry of *custom* will result in an analysis based on a user-defined transform. The preferred transformation and its inverse must be defined using columns C, D, E, and F of the respective transform row. Column C would contain a representative value of *a* or $\hat{a}$. Column D would be the Excel equation defining the custom transform on the value in column C. Similarly, Column E would contain a representative transformed value, and column F would be the Excel equation defining the inverse transformation.

When an analysis is performed using transformations, POD is first calculated in terms of the transformed crack sizes and then converted to the measurement units of the recordings. All parameter estimates on the **Result** sheet are expressed in terms of the transformed variables but a_{90} and $a_{90/95}$ values are in the units of input size measurements. The **Fit** plot sheet for visually judging the goodness of fit is also expressed in terms of the transformed variables. The **Residuals** plot sheet is presented in terms of crack size versus the transformed $\hat{a}$ values. The **ahat versus *a*** plot sheet presents the fit based on the transformed data superimposed on the log-log plot of $\hat{a}$ versus *a*. Threshold plots and POD(*a*) plots with confidence bounds are presented in terms of the original input measurement units.

3.2.2 Tests of Assumptions for $\hat{a}$ versus a Analysis

Because the $\hat{a}$ versus a analysis depends on the assumptions of linearity, equal variance, and normality, statistical tests of the significance of the assumptions were programmed into the analysis. The results of the three significance tests are included on the **Results** worksheet of the workbook. For each test, the calculated test statistic is listed and its statistical significance is evaluated in terms of the probability, P, of obtaining the test statistic or worse if the assumption is true. Small values of P, such as those less than 0.05, are indicative that the assumption is not valid. Values of P that are greater than 0.10 are indicative that the assumption is reasonably valid. Descriptions of the programmed significance tests appear in the following subparagraphs. If different tests of significance are desired, they can be performed in the Excel workbook.

The $\hat{a}$ versus a analysis is performed using the averages of the $\hat{a}$ responses from each of the cracks of the specimen set. Accordingly, the three assumptions are tested using the differences between the average $\hat{a}$ values and the estimated linear fit. Such differences are called the residuals. Only the average $\hat{a}$ values that are between the censoring limits of the minimum (signal threshold) and maximum (saturation) are included in the hypothesis tests. Average $\hat{a}$ values when at least one, but not all, $\hat{a}$ values are uncensored are calculated using best linear unbiased estimators as described in Lawless [19].

3.2.2.1 Linearity

The linearity of a fit is evaluated using the pure error lack of fit test [20]. In the pure error test, the sum of squares of all residuals is considered to comprise the part due to cracks of the same size and the part due to lack of fit from the linear model. If the model is nonlinear, the sum of squares due to the lack of fit will contain biases that will inflate the lack of fit variance. The change in variance can be seen in the example residual plot of Figure 19. It is easy to see that the scatter of residuals about the linear fit is larger than the scatter of the residuals for cracks of the same size. When estimating the residual variance for cracks of the same size, the mean is in the middle of the ln $\hat{a}$ values for cracks of that size. When estimating the variance of the residuals from the linear fit, the mean is the line at each crack size and there is a large bias at both the large and small crack sizes.

The pure error test is an F test that compares the variance of the residuals from the lack of fit to the pooled variance of the residuals from cracks of the same size. The F statistic is a ratio of variances. Under the null hypothesis that the variance of lack of fit residuals is equal to the variance of residuals for cracks of the same size (i.e., the pure error), large values of F are indicative that the biases associated with the fit are significant. The POD program automatically performs the pure error lack of fit test whenever a fit is generated. The probability of obtaining an F value as large or larger than that calculated from the data is included in the output on the **Results** worksheet.

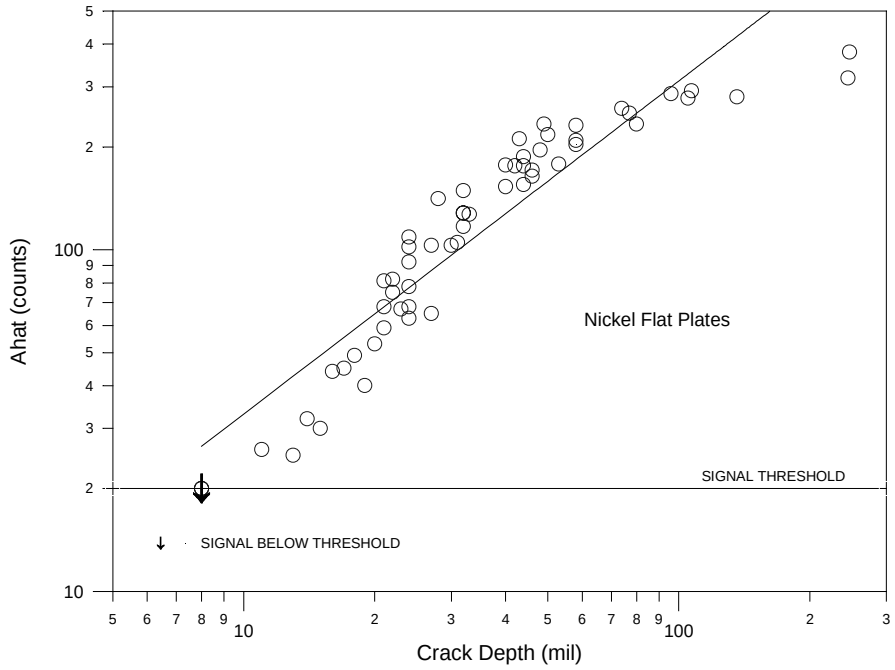


Figure 19. Example $\hat{a}$ versus a Plot Exhibiting Smaller Residual Variance for Cracks of the Same Size as Compared to Variance of All Residuals

Nonlinearity is quite easy to detect visually when there is only one independent variable and the visual test may well be more sensitive to nonlinearity than the formal hypothesis test. When a visual judgement and the hypothesis test do not agree, the analyst must consider that the fit in the crack size range of primary interest. The goal of the analysis is developing a fair description of the scatter in inspection responses about the estimated mean. The plot of the residuals versus crack size aids in the evaluation of the fit. This residual plot will be generated by POD at the request of the analyst.

3.2.2.2 Homogeneity of Variance

Homogeneity of variance (the standard deviation of residuals being independent of crack size) is tested using a standard Bartlett's test [21]. Experience has shown that when changes in the scatter of responses are detected, the changes are most often caused by increased scatter at the small crack sizes. On occasion, an increase in scatter has been detected that is apparently associated with the signal response approaching a nonsaturation maximum. Based on this experience, the cracks are partitioned into small and large crack regions and Bartlett's test is used to compare the variances of the residuals from the two regions. Bartlett's test statistic has a χ^2 distribution and large values of χ^2 are inconsistent with the assumption of equal variances. The probability of obtaining a χ^2 value as large or larger than that calculated from the data is included in the output on the **Results** worksheet.

Homogeneity of variance is not always readily detectable by visually examining the residuals. The plot of residuals versus crack size, however, is an aid in judging the degree

of scatter and the crack sizes at which the scatter is lesser or greater. If the primary focus of a_{90} values is in the range of the smaller standard deviation of residuals, ignoring the difference will lead to more a conservative estimate of a_{90}. See subsection 2.2.2.2 for a discussion of this topic.

3.2.2.3 Normality

The test of normality of the residuals from the fit is performed using the Anderson-Darling test [22]. The Anderson-Darling test is based on a quadratic average of the difference between the sample cumulative distribution of residuals and a normal distribution with parameters calculated from the residuals. Large values of the test statistic, A*, are indicative of non-normality. The probability of obtaining an A* value as large or larger than that calculated from the data is included in the output on the **Results** worksheet.

When the linearity and equal variance hypotheses have been reasonable, normality has usually been accepted in the $\hat{a}$ versus a data from the ECIS system. When normality was rejected, one outlying $\hat{a}$ value would usually be the cause of the rejection. On most such occasions, when the crack was inspected at different times and with different probes or stations, the resulting $\hat{a}$ value tended to be an outlier from the other cracks in the specimen set. The outlying response may well have been due to mis-sizing the crack (see subsection 2.1.3). In all such cases, the added variation due to the outliers was included in the estimate of the scatter about the fit but the lack of normality was ignored in the analysis. If normality is reasonable, this approach will yield slightly conservative a_{90} values. Since POD estimates the parameters only for a normal distribution of responses about the mean fit, when normality is rejected, the POD results are still reported. Special note should be made of the possible non-normality.

3.2.3 Alternate Values of POD and Confidence Level

It has been customary to characterize inspection capability in terms of a 95 percent confidence bound on the 90 percent detectable crack size. The updated program has the capability to calculate a 90, 95, or 99 percent confidence bound on the arbitrarily specified POD value. The 95 percent confidence bound on the 90 percent detectable crack size, $a_{90/95}$ is the default calculation.

3.3 Output of POD

A full POD workbook comprises 11 named spreadsheets. The **Info** and **Data** spreadsheets contain the required input data for an analysis and are discussed in subsection 3.1. The other nine named spreadsheets contain the output of an analysis and comprise a maximum of four data type sheets and five figures. The first analysis of a data set will produce two information or data type spreadsheets. The figures must be specifically requested through the POD window, but once requested, are automatically generated for additional recalculations. Two of the figures require the calculation of the two additional

data type spreadsheets. All of these output spreadsheets are described in the following paragraphs. Only the pertinent output sheets are available in a pass/fail analysis.

3.3.1 Results Sheet

The **Results** sheet contains a summary of the results of the analysis and is different for the $\hat{a}$ versus a and pass/fail analyses. Consider first the **Results** sheet for an $\hat{a}$ versus a analysis for which an example is presented as Figure 20. The top lines repeat all of the identifying information from the *Title* and *Subtitle* lines of the **Info** sheet. Next are the range of crack sizes, the number of cracks in the analysis, the number of censored recordings, and the inspection titles. In the example sheet(Figure 20), there are two inspections per crack. Both responses related to four of the cracks had less than the minimum (signal threshold), and one crack had one response above and one below the minimum. None of the cracks had $\hat{a}$ values above the maximum (saturation).

The analysis type and model of the analysis are followed by the parameter estimates and standard errors of the $\hat{a}$ versus a fit parameters. The standard errors are the standard deviation of the estimates of the parameters and indicate the degree of precision of the estimates. The *Residual Error* is the standard deviation of the differences between the average $\hat{a}$ values and the linear fit (σ_{δ^*} of Equation 10). The repeatability error is the pooled standard deviation of the repeated $\hat{a}$ values for each crack (σ_p of equation 10). The results of the hypothesis tests for model fit are presented in terms of the calculated test statistic and the significance level of the test. High values of P indicate that the data are compatible with the assumption.

The POD parameters are summarized only in terms of *Sigma* and a_{50}, a_{90}, and $a_{90/95}$ values for the *POD Threshold* on the **Info** sheet. The parameter μ depends on the decision threshold, $\mu = \ln(a_{50})$. The variance-covariance matrix for the estimates of μ and σ are contained in a different worksheet.

Note that all information concerning the parameters of the fit are expressed in terms of the transformed variables. The a_{50}, a_{90}, and $a_{90/95}$ values are in the crack size units.

An example **Results** sheet for a pass/fail analysis is presented in Figure 21. The top lines repeat all of the identifying information from the *Title* and *Subtitle* lines of the **Info** sheet. Next are the range of crack sizes, the number of unique crack sizes, and the total number of cracks in the analysis. This example has one inspection that is called Ins 1. The analysis and model are identified followed by the estimates of the POD(a) parameters. Estimates of a_{50}, a_{90}, and $a_{90/95}$ are listed along with the variance-covariance matrix of the parameter estimates. The variance-covariance values are used in equation 30 to obtain the confidence limits for $a_{p/95}$.

Titanium Small Bolt Holes
Example line 2
Example line 3
Example line 4
Example line 5
Example line 6

Flaw Depth Range: 3 to 17 mil

				some	all
Cracks analyzed:	42				
Signal Minimum:	50	counts	below	1	4
			between		37

Inspections: Ins 1 Ins 2

"ahat vs. a" Analysis **Version:** POD 3.0

Model: ln(ahat) = B0 + B1*ln(a)

Parameter	Estimate	Std. Error
Intercept (B0)	0.856657	0.371049
Slope (B1)	1.829921	0.155633
Residual Error	0.300438	0.034897
Repeatability Error	0.102684	

Tests of Assumptions

			P, if true
Normality: Anderson-Darling	A* =	0.414763	P > 0.1
Equal Variance: Bartlett	χ^2 =	0.057012	P > 0.1
Lack of fit: Pure Error (df=27)	F =	1.620293	P > 0.1

POD Model Parameters

Sigma 0.168907

Inspection Threshold	a50	a90	a90/95
250	12.79665	15.89052	18.43548

Figure 20. Example POD **Results** Sheet for *â* versus *a* Analysis

3.3.2 Residuals Sheet

The **Residuals** sheet for an *â* versus *a* analysis contains four tables that are used in generating plots. These are the residuals, *â* versus *a*, fit, and min/max tables. The residual table lists only the cracks used in the analysis and includes the crack sizes, the average *â* for each crack, the logs of crack size and average *â*, and the differences (residuals) between the average and predicted *â* values. The *â* versus *a* table lists all of the cracks from the **Data** sheet and contains the crack sizes, the individual *â* values from all inspections, and the predicted *â* for the crack size. The fit table is a small table of the straight line fit for the

data. The min/max table is used to plot the minimum and maximum $\hat{a}$ value over the range of data.

The **Residuals** sheet for pass/fail data comprise the observed proportion of the cracks of each size that were detected, the POD(a) estimate for each crack size, and the difference between the observed and estimated POD.

Ultrasonic Surface Wave Inspection

Flaw Depth Range:	2.5	to	27.8116	mil
# of unique cracks:	46	of	144	valid cracks

Inspections: Ins 1

Pass/Fail Analysis

Model: log normal

POD Parameters

Mu-hat	1.557202
Sigma-hat	0.599553

Percentile Estimates			Estimated Covariance Matrix		
a50	a90/50	a90/95	V11	V12	V22
4.745527	10.23241	15.18824	0.010588	-0.0061	0.01147

Figure 21. Example POD **Results** Sheet for Pass/Fail Analysis

3.3.3 Threshold Data Sheet

The **Threshold Data** sheet is a table that contains a_{50}, a_{90}, $a_{90/95}$, V_{11}, V_{12}, and V_{22} for ranges of thresholds that are specified for the analysis on the **Info** page. The program will insert nine equally spaced threshold values between those listed on the **Info** sheet. There is no **Threshold Data** sheet in a pass/fail analysis.

3.3.4 POD Data Sheet

The **POD Data** sheet comprises three columns that contain the array of crack sizes, the estimated POD(a) function, and the confidence bound for POD(a).

3.3.5 Ahat vs a Sheet

The **Ahat vs a** sheet contains a plot of $\hat{a}$ versus a for all of the inspections of all of the cracks. An example of this plot for the data from Figures 17 and 20 is presented in Figure

22. This plot aids in the selection of the range of cracks or the model formulation to be used in the POD(a) analysis. It also can identify individual inspection results that do not agree with other inspections of the same crack. There is no **Ahat vs a** sheet in a pass/fail analysis.

3.3.6 Fit Plot Sheet

The **Fit Plot** sheet for $\hat{a}$ versus a data contains a plot of the average $\hat{a}$ for each crack versus the crack size with a superimposed straight line fit obtained from the analysis. Figure 23 presents an example fit plot for the data of Figures20 and 22. The fit plot provides for easy visual inspection of the goodness of fit and can be used to choose crack size regions for which the relation might be more linear.

The **Fit Plot** sheet for a pass/Fail analysis is a plot showing the POD(a) fit on the observed detection probabilities. Figure 24 is an example pass/fail **Fit Plot** for the **Results** sheet of Figure 21. A subjective judgement of goodness of fit can be made from this plot.

3.3.7 Residual Plot Sheet

The residual plot presents the difference between average and predicted $\hat{a}$ as a function of the size of the crack. An example residual plot is presented in Figure 25 for the fit of Figure 22. The residual plot for $\hat{a}$ versus a data aids in identifying crack size regions for which the fit may not be linear or for which the scatter in residuals is changing. The residual plot is also useful in identifying the outlying data points that may be affecting the tests of hypotheses.

3.3.8 Threshold Plot Sheet

The threshold plot presents the estimates of a_{90} and $a_{90/95}$ as functions of the decision threshold. The threshold plot for the data of Figures 20, 22, 23, and 25 is presented in Figure 26. This plot has become the most useful characterization of NDE capability for $\hat{a}$ versus a data because a demonstration of capability is often used for different target a_{90} values. Further, thresholds in automated systems often need to be adjusted and the threshold plots readily yield the a_{90} values that would result for different choices. There is no **Threshold Plot** sheet in pass/fail analysis.

3.3.9 POD Sheet

The **POD** sheet contains the POD(a) function and the confidence bound that has been a prime objective in the characterization NDE capability. Figure 27 presents an example plot of the "**POD** sheet for the data of Figures 20, 22, 23, and 25. In an $\hat{a}$ versus a data analysis, the threshold of the POD(a) function is that identified as the *POD Threshold* on the **Info** sheet.

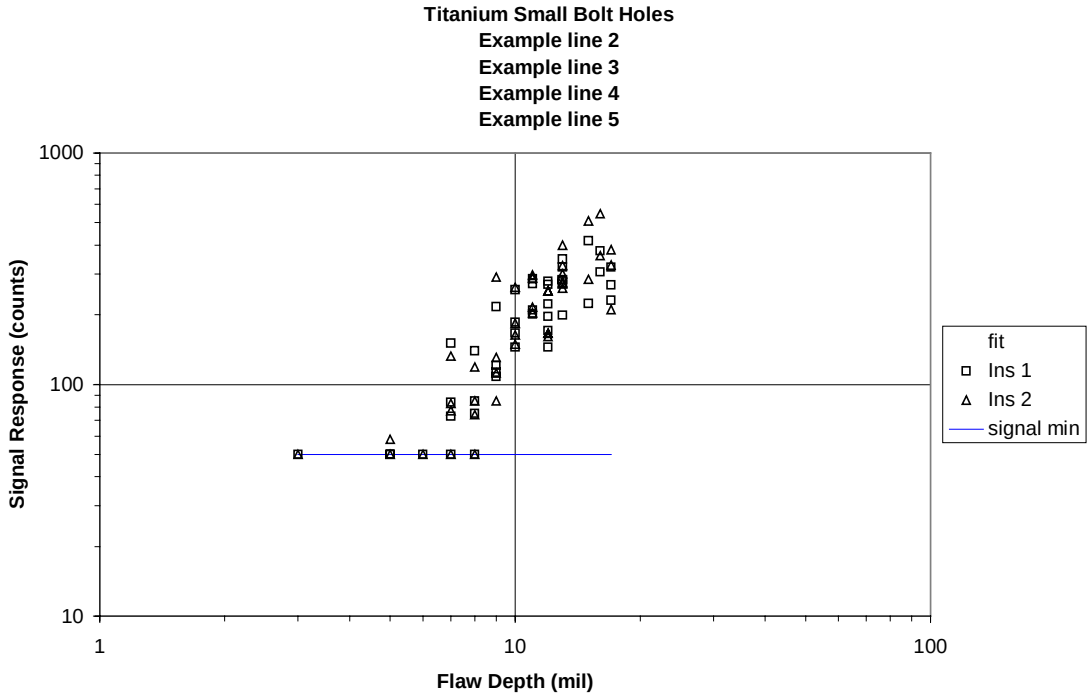


Figure 22. Example **Ahat versus a** Sheet

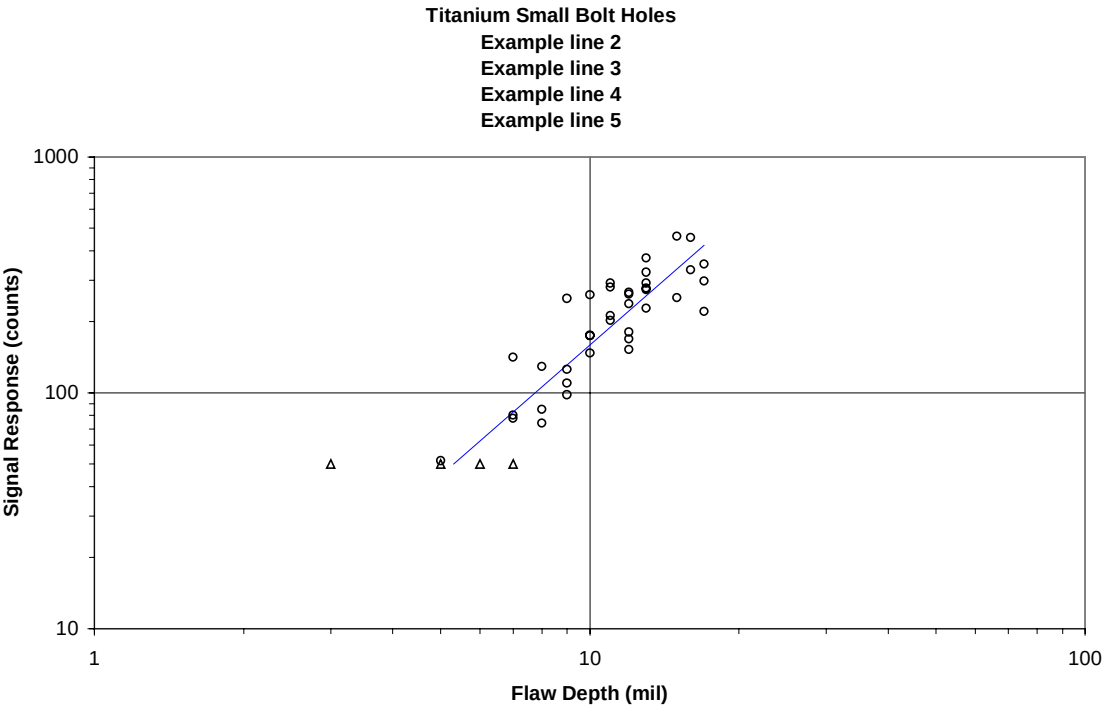


Figure 23. Example **Fit Plot** Sheet for $\hat{a}$ versus a Data

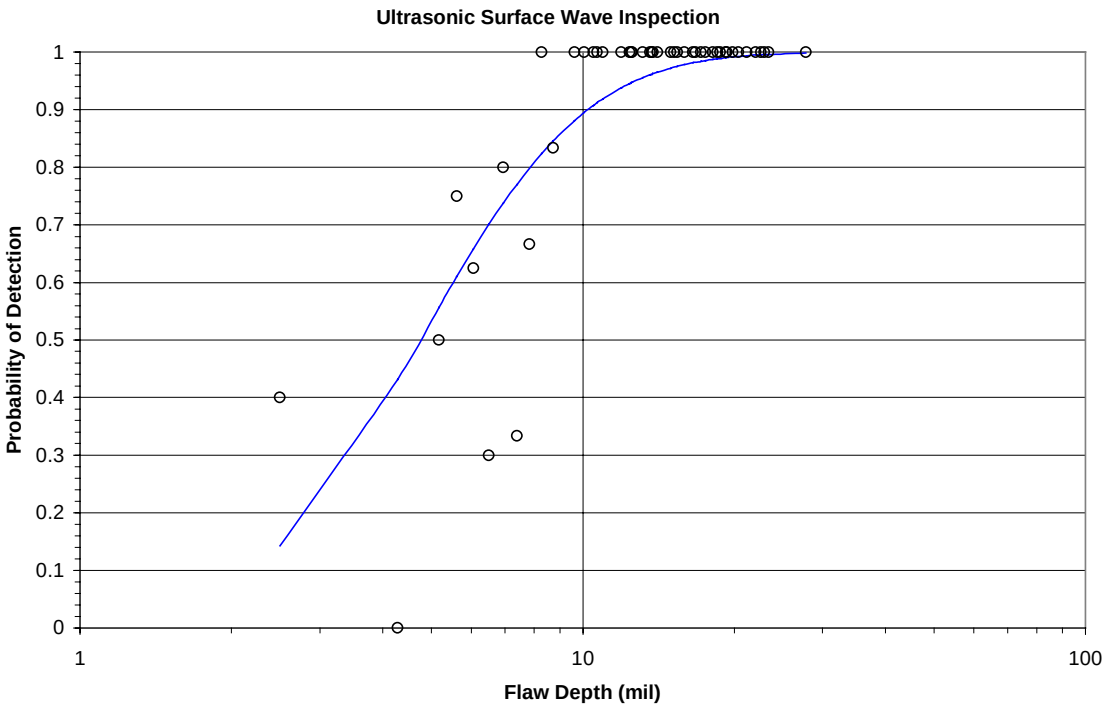


Figure 24. Example **Fit Plot** Sheet for Pass/Fail Data

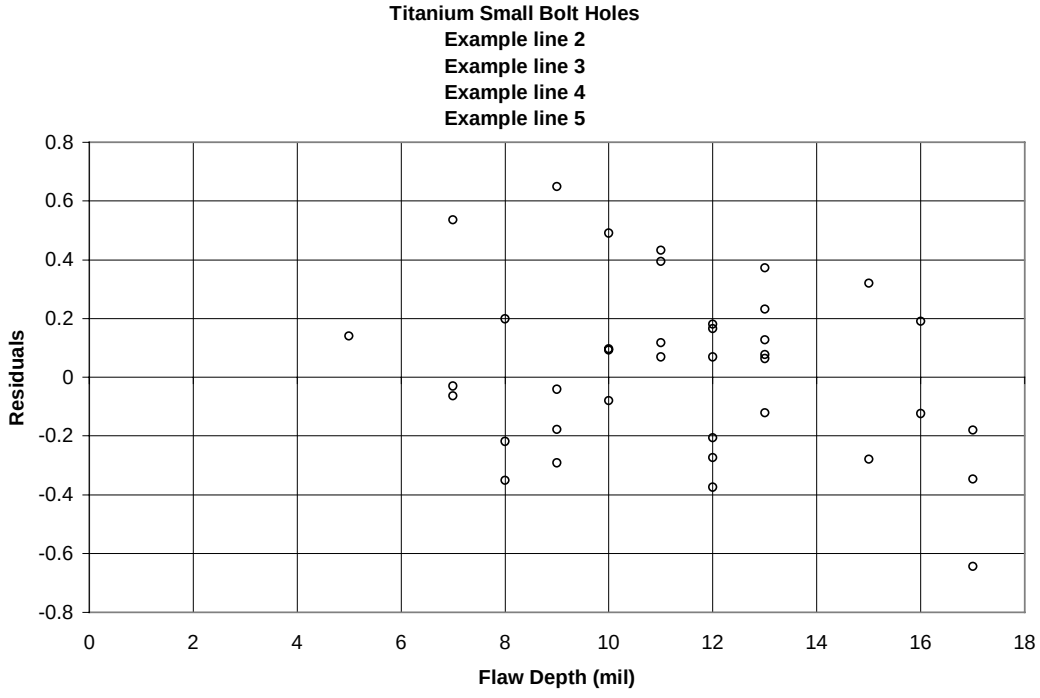


Figure 25. Example **Residual Plot** Sheet

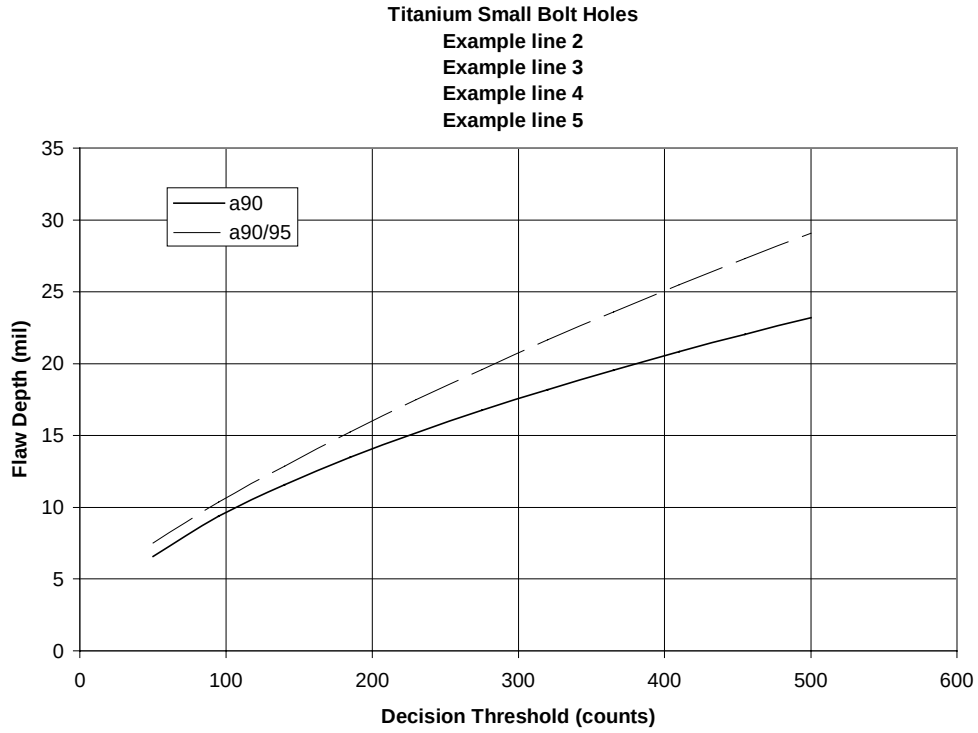


Figure 26. Example **Threshold Plot** Sheet

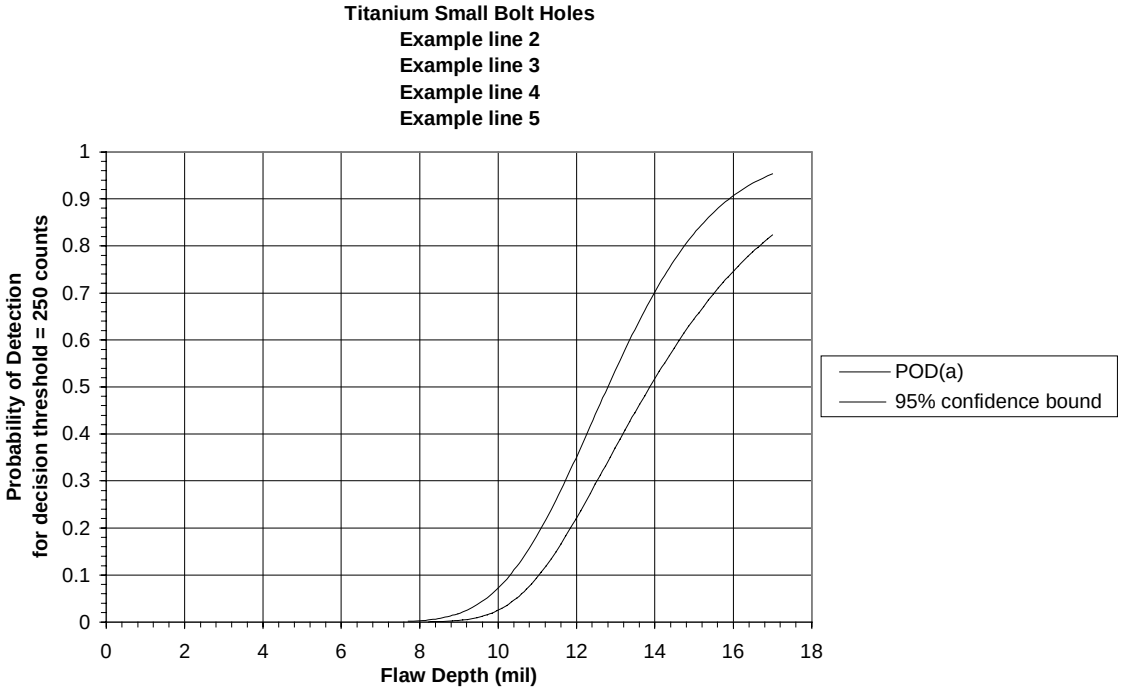


Figure 27. Example **POD** Sheet

Section 4
RFC/ENSIP ECIS Evaluations

Several modifications were developed for the NDE ECIS as part of a system upgrade. The changes include: a) the eddy current instrument, b) the PC-based RFC station computer and operating system, c) the controller, and d) the scanner. In addition, a new method for calibrating the system was demonstrated. A prime criterion for the acceptance of these changes was that they would be "drop in." That is, the changes could be implemented using the previously established $\hat{a}$ detection thresholds, which are selected to yield a reliably detected crack size of a_{90} in the RFC/ENSIP application.

This section presents the analysis of the data collected to demonstrate the drop-in compatibility of each of the individual components and the data collected from the complete integrated system. The validation studies reported herein deal only with the inspections of the reliability specimens. Other validation demonstrations are discussed in the reports describing the individual system modifications. Since the calibration methodology is not being implemented at this time, the compatibility of the immediate modifications is presented first.

4.1 Eddy Current Instrument Validation

The NDT25L eddy current instrument of the ECIS is being upgraded by the substitution of the US500L instrument [23]. To demonstrate the drop-in compatibility of the new instrument, Ti-6246, 0.155-inch bolt hole and Waspaloy flat plate reliability specimen sets were inspected at both 2 and 6 MHz. The 2-MHz bolt hole, 2-MHz flat plate, and 6-MHz flat plate inspections were repeated using two different probes. The inspections were repeated using three different probes in the 6-MHz bolt hole inspections. In each of the four cases, the same probes were used in both instruments. These inspections of the reliability specimens were conducted as a complete factorial experiment. Comparisons of the sets of results from the two instrument types were made by directly comparing the magnitudes of the $\hat{a}$ signals from the two instruments. If the $\hat{a}$ values are not statistically different, POD(a) evaluations from the two instruments would be equivalent, and previously applied threshold would yield the same a_{90} values.

Figures 28 through 31 present log-log plots of average $\hat{a}$ for each crack for the NDT25L versus the US500 instruments. If the averages for each crack were exactly equal, all data points would lie on the 45 degree line. The comparisons are shown on a log-log plot because the POD(a) function is derived from a log-log plot of crack size versus $\hat{a}$. The scatter exhibited in these plots is within that often seen as a result of a probe change between inspections of the same cracks.

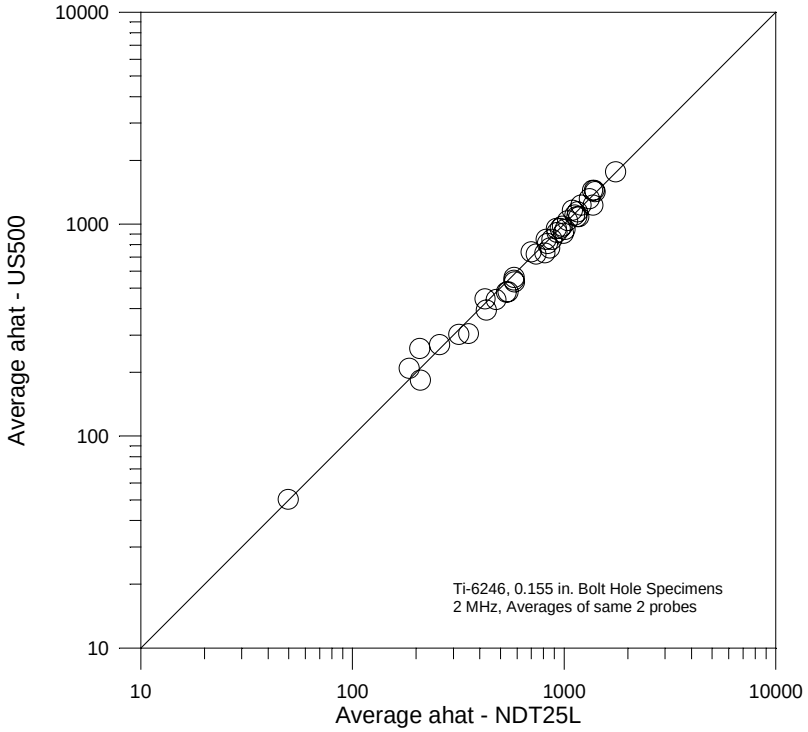


Figure 28. Comparison of Average $\hat{a}$ for Ti-6246 Bolt Hole Specimens, 2 MHz

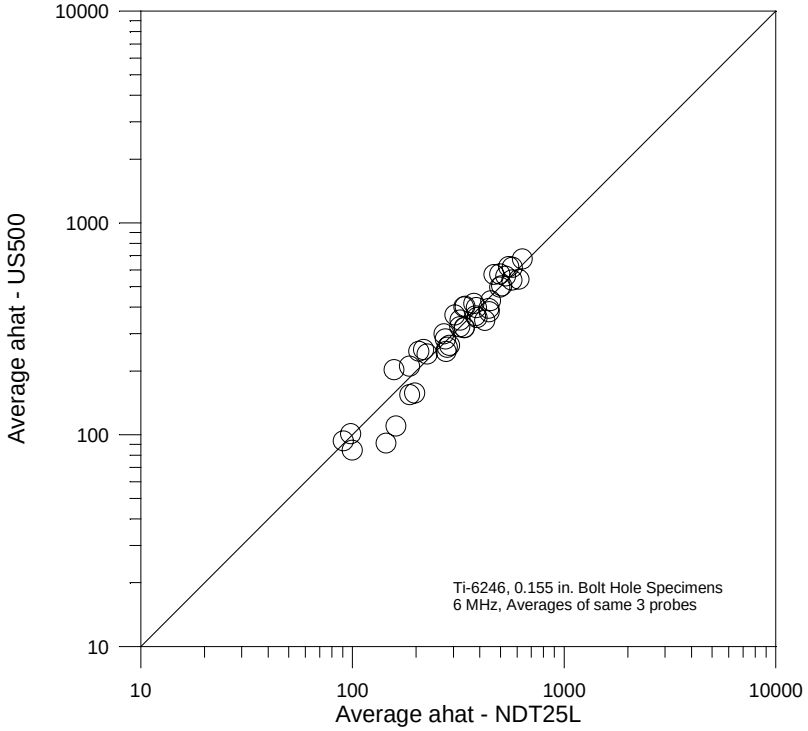


Figure 29. Comparison of Average $\hat{a}$ for Ti-6246 Bolt Hole Specimens, 6 MHz

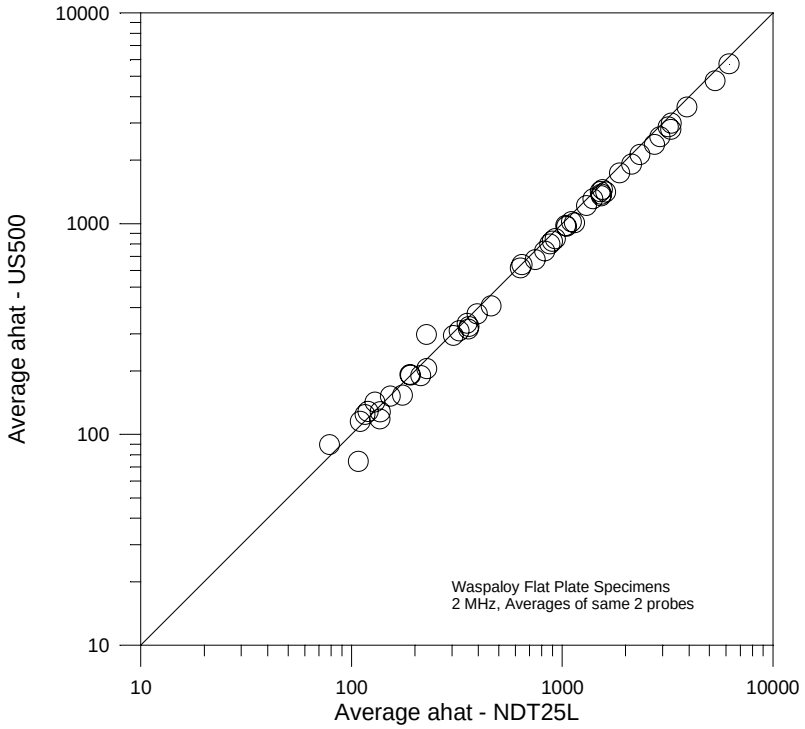


Figure 30. Comparison of Average $\hat{a}$ for Waspaloy Flat Plate Specimens, 2 MHz

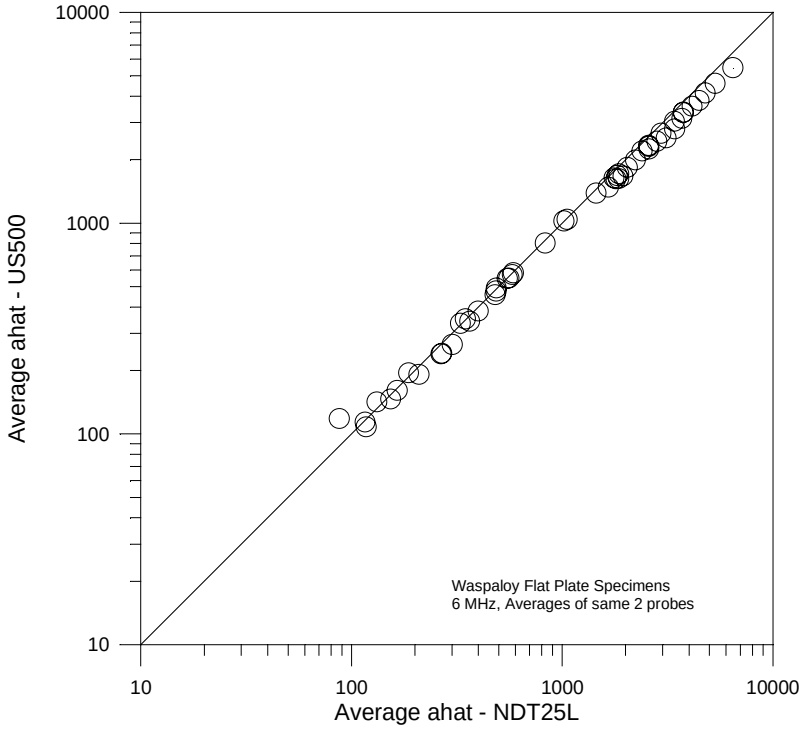


Figure 31. Comparison of Average $\hat{a}$ for Waspaloy Flat Plate Specimens, 6 MHz

To further investigate potential differences, for each specimen type by frequency combination, an analysis of covariance was performed. The analysis of covariance first accounts for the differences in $\hat{a}$ magnitudes that are due to crack size and then compares the effects of instrument, probe/calibration, and the interaction. The probe and calibration effect must be considered jointly as a probe change necessitates a recalibration. In all four experiments, the effect of eddy current instrument was not statistically significant. In three of the four experiments, the probe/calibration effect was significant. These results indicate that the differences in the $\hat{a}$ response from the two instruments is less than the effect of changing and recalibrating probes for the same instrument.

To demonstrate that the US500 eddy current instrument would yield the same decision thresholds as the NDT25L, plots of a_{90} versus a_{dec} were generated for each of the inspections on the Ti-6246 bolt hole and Waspaloy flat plate specimen sets. These comparisons are presented in Figures 32 through 35. Since it is common for a_{90} values to differ by about a mil or so due to changes in probes and recalibrations alone, the minor differences that appear in these plots are negligible.

Assuming that the results of the above evaluations are typical of that which would be expected in other combinations of material and geometry, it can be concluded that the NDT25L and US500 instruments produce equivalent $\hat{a}$ values and, hence, equivalent POD(a) and a_{90} values for the same detection thresholds.

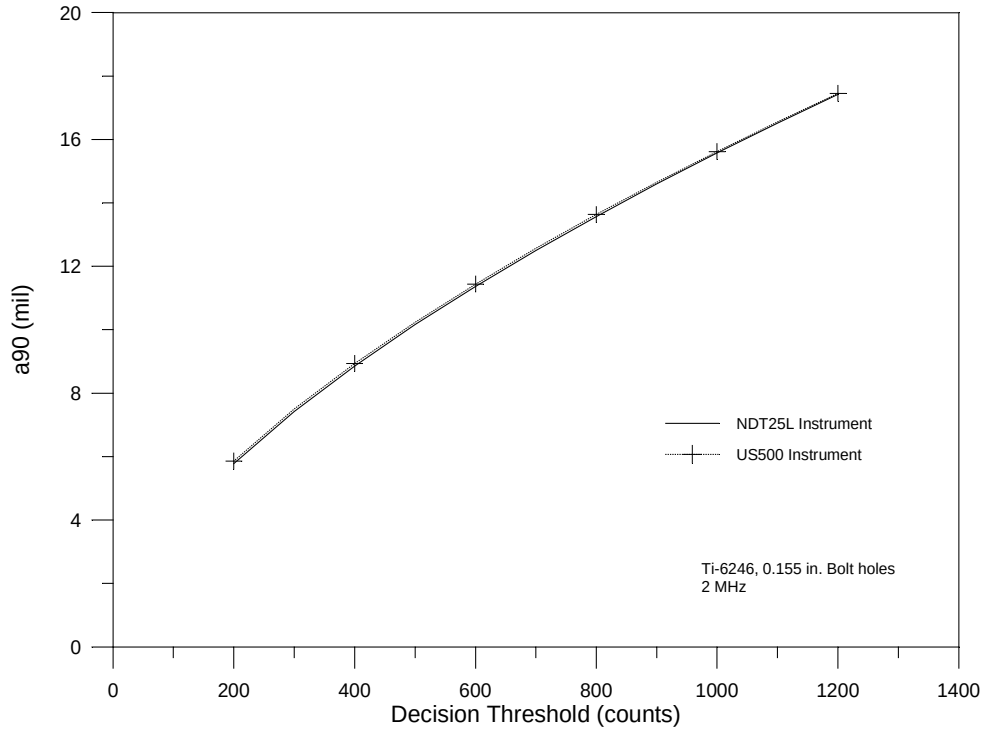


Figure 32. a_{90} Threshold Comparisons for Eddy Current Instruments – Ti-6246, 0.155" Bolt Holes, 2 MHz

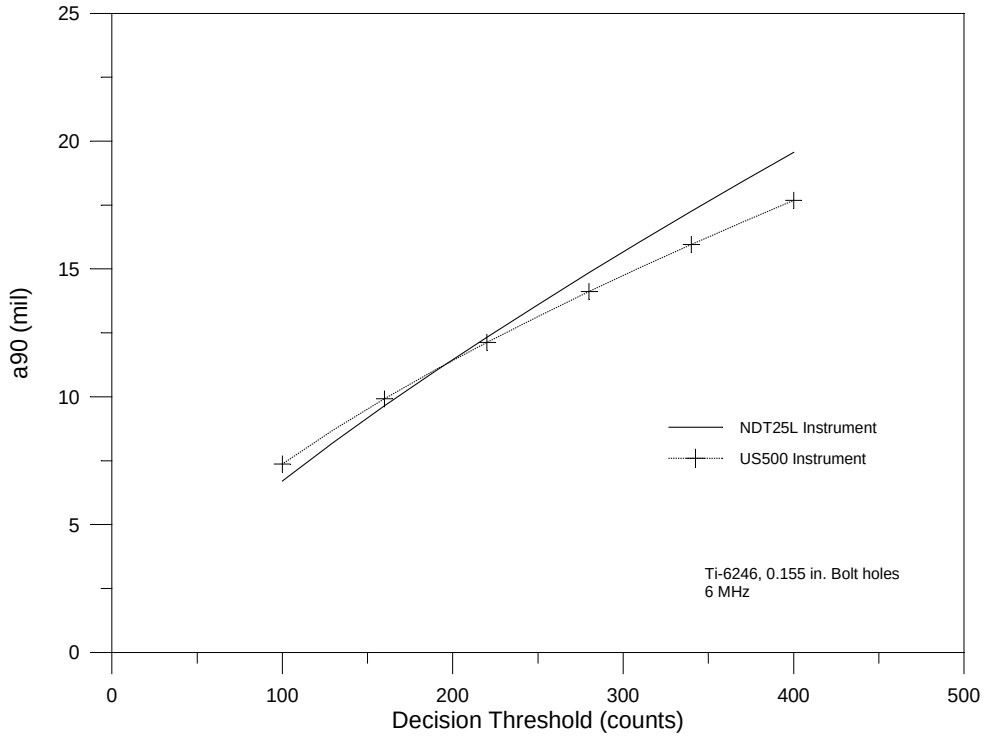


Figure 33. a_{90} Threshold Comparisons for Eddy Current Instruments – Ti-6246, 0.155" Bolt Holes, 6 MHz

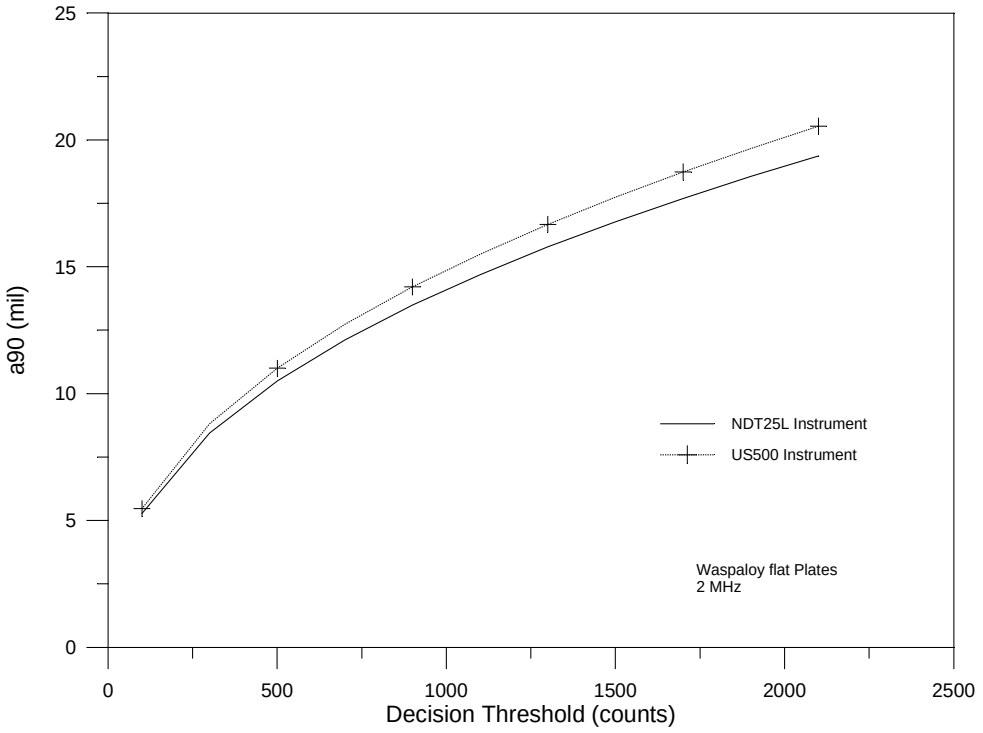


Figure 34. a_{90} Threshold Comparisons for Eddy Current Instruments – Waspaloy Flat Plates, 2 MHz

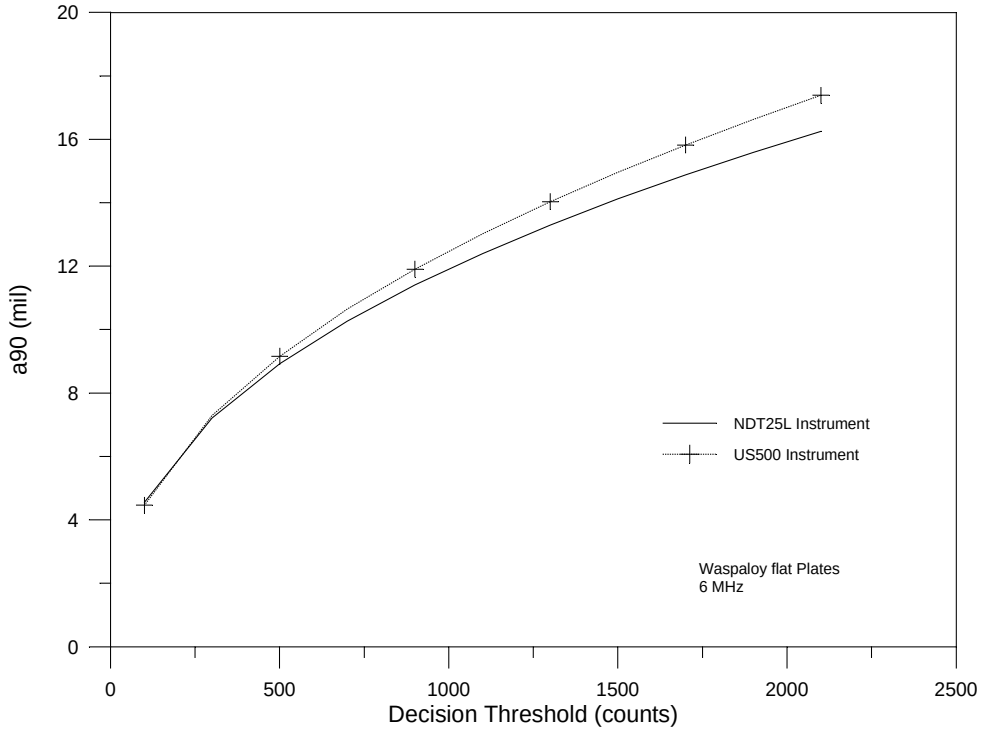


Figure 35. a_{90} Threshold Comparisons for Eddy Current Instruments – Waspaloy Flat Plates, 6 MHz

4.2 Station Computer Validation

A major modification planned for the ECIS is the replacement of the original Intel station computer with a PC-based computer with an NT operating system [24]. As part of the validation of the drop-in compatibility of the replacement computer, 17 scan plans were exercised on reliability specimen sets with both the original and replacement computers. Table 4 presents the scan plans, the number of cracks inspected, the probes used in each inspection and the number of repeats of each inspection.

Table 4. Specimen Test Matrix for PC Station Computer Validation

		PC Station Computer		Original Station Computer	
Scanplan	**# Cracks**	**Probe #'s**	**Insp/probe**	**Probe #'s**	**Insp/probe**
dhole-rel	23	51294	2	51294	1
dt-rel	64	932095	2	932095	2
r88-scal-20	107	13956	1	13956, 15310, 15311	1
r95-bh-50	75	932095, 932228	1	932095, 932228	2 and 1
r95-fp-40t	44	14014, 14015, 14016	1	14014, 14016	2 and 1
in100-fp1	6	275-1, 1-691	2 and 1	275-1, 1-691	1
ti6246-fp1	5	275-1, 1-691	1	1-691	1
in100bh	6	1093024, 306	1	s/n 40	1
wasp-bh	5	1093024, 40	1	1093024, 40	1
ti17-bh2-ts	75	932095	2	932095	1
ti17-dtrel-d20	69	913	2	932095	2
ti6246-as-20	14	13956	1	13956, 13567, 13957	1
ti6246-bse-20	10	900	1	900, 901, 902	1
ti6246-es-20	28	13956	1	13568, 13567, 705-902	1
ti6246-ss-20	17	?	1	13567, 13568, 902	1
ti6246-ss-b	36	902	1	902, 903, 904	1
ti17-dtrel	34	913	2	913	1

The data from this station computer evaluation were not collected in accordance with a consistent design plan. Comparisons of the inspection results from the two station computers were made on the basis of comparable $\hat{a}$ values and calculated a_{90} values for those data sets for which sufficient data were available to perform a POD(a) analysis. Table 5 summarizes these comparisons.

Table 5. Summary Statistics from PC Computer Validation

			***a90* at min threshold**		***a90* at max threshold**	
Scanplan	**# Cracks**	**$\hat{a}$ Ratio**	**PC Station**	**Orig Station**	**PC Station**	**Orig Station**
dhole-rel	23	1.08	8.8	8.9	23.6	24.4
dt-rel	64	1.38	12.9	14.6	49.4	53.3
R88-scal-20	107	1.02	10.5	10.3	19.9	19.4
R95-bh-50	75	0.97	11.4	11.1	24.1	23.8
R95-fp-40t	44	0.96	11.3	10.8	26.4	25.5
Ti17-bh2-ts	75	0.98	23.1	22.2	64.0	64.1
Ti17-dtrel-d20	69	1.20	9.6	10.2	63.6	70.3
Ti-6246-es-20	28	1.05	7.0	7.2	20.1	22.1
Ti-6246-ss-b	36	1.25	4.1	5.2	17.8	20.6
Ti17-dtrel	34	0.97	20.3	20.1	31.2	30.7

For 7 of the 10 specimen sets, the data obtained from the 2 computers agreed within the variation commonly seen in POD(*a*) evaluations from the ECIS when multiple probes are used. For three of the sets, the inspections using the PC station computer produced $\hat{a}$ values that were significantly larger than those from the original computer when comparing data from the same probes. The Ti-6246-ss-b produced $\hat{a}$ values that were 25 percent larger. However, when a different probe was used the with the original computer, the $\hat{a}$ readings still differed by 20 percent. The Ti17-dtrel-d20 scan plan produced $\hat{a}$ values from the PC station that were 20 percent larger than those from the original computer. Different probes were used to collect these two sets of data, and the difference might well be a probe to probe difference. The dt-rel scan plan tests resulted in a 25 percent average difference and no cause for this discrepancy could be determined. Figure 36 presents a plot of average log $\hat{a}$ values from the PC station plotted against the log $\hat{a}$ values from the original station. As can be seen in the figure, the difference at most of the crack sites is a reasonably constant ratio. The causes for the four or so aberrant data points were not determined. Since a multiplier shift in log response should be accounted for by the calibration, it was concluded that the differences in response were more likely due to some cause other than the station computer.

Assuming that the results from this comparative study of data from the two station computers are representative of all inspections, it was concluded that the PC station computer with the NT operating system is a drop-in substitution for the original Intel station computer.

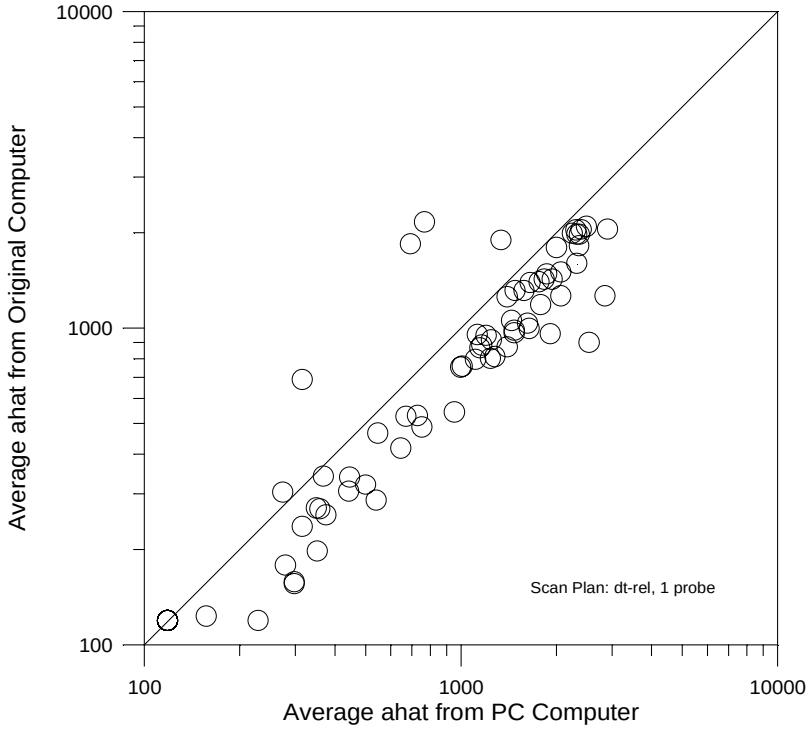


Figure 36. Comparison of $\hat{a}$ Values from the dt-rel Scan Plan

4.3 Robotic Controller Validation

An American Robotics controller will replace the original in the ECIS [25]. In order to test the drop-in compatibility, data from the original controller and the American Robotics controller from four reliability specimen sets were compared. The specimen sets used in these demonstration tests were the Waspaloy flat plate, IN718 bolt hole, and Ti-6246 elongated scallop specimens from the F100-PW-229 program and the Ti-6246 broach slot (mid thickness) from the F100-GE-220 program. All four specimen sets were inspected using the American Robotics controller. The Waspaloy flat plate, IN718 bolt hole and Ti-6246 elongated scallop data were then compared with the results obtained during the capability demonstration for the F100-PW-229 engine that had been conducted two years previously. The Ti-6246 broach slot specimens were inspected with the original controller at the same time as the inspections with the American Robotics controller. The cracks in the specimen sets were inspected twice with the American Robotics unit and two or three times with the original unit. The probes and ECIS stations were not necessarily the same for the comparison within specimen sets.

Figures 37 through 40 present the comparisons between the average $\hat{a}$ values from the American Robotic and original units. The differences in $\hat{a}$ values between the two controllers would have produced insignificant changes in decision thresholds, Table 6. The maximum difference in a_{90} values over the valid range of analysis in any of the four data sets was 1.5 mil. This largest discrepancy occurred in the Waspaloy flat plate specimen set for which the difference might be attributable to a calibration or ECIS station difference.

Table 6. Comparison of a_{90} Values from American Robotics and Original Controller Units

	WASP FP		IN718 BH		Ti-6246 BSM		Ti-6246 ES	
â thr	**AR**	**Orig**	**AR**	**Orig**	**AR**	**Orig**	**AR**	**Orig**
50			5.1	5.6				
75			7.2	7.8	9.2	8.9		
100			9.1	9.8	10.7	10.3		
125			11.0	11.7	12.0	11.5		
150			12.7	13.6	13.2	12.6		
175			14.5	15.3	14.3	13.7		
200			16.2	17.1	15.4	14.6		
250	7.2	8.7			17.2	16.4	4.4	4.9
500	10.2	11.7					6.3	6.8
750	12.5	13.9					7.9	8.2
1000	14.4	15.7					9.2	9.3
1250	16.1	17.3					10.3	10.3
1500	17.7	18.7					11.4	11.2
1750	19.1	20.0						
2000							11.1	9.9
2250							12.3	11.1
2500							13.4	12.3
2750							14.5	13.5
3000							15.5	14.7
3500							17.7	17.0
4000							19.7	19.4
5000							23.7	24.0

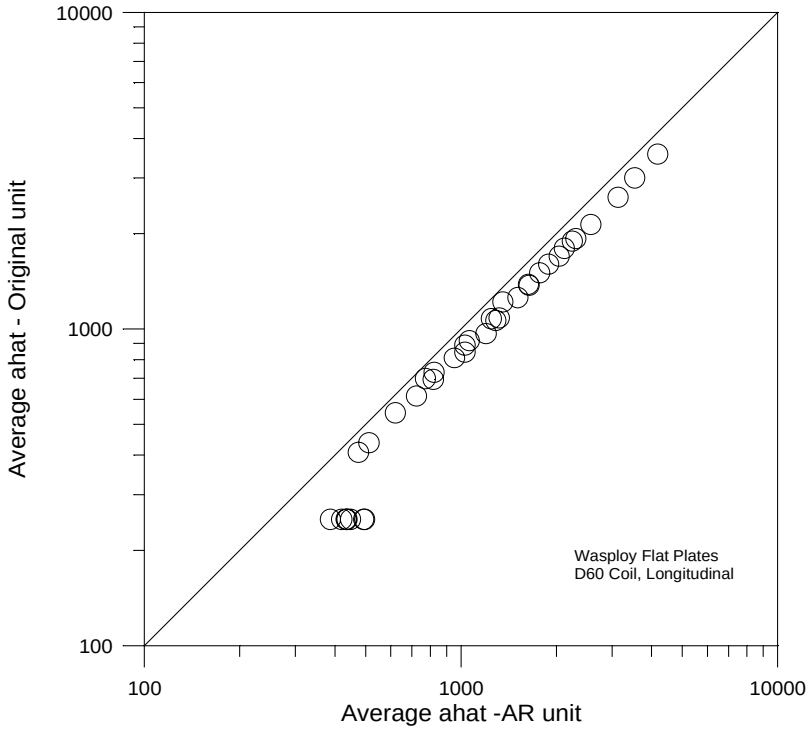


Figure 37. Comparison of $\hat{a}$ Values from Controllers – Waspaloy Flat Plates

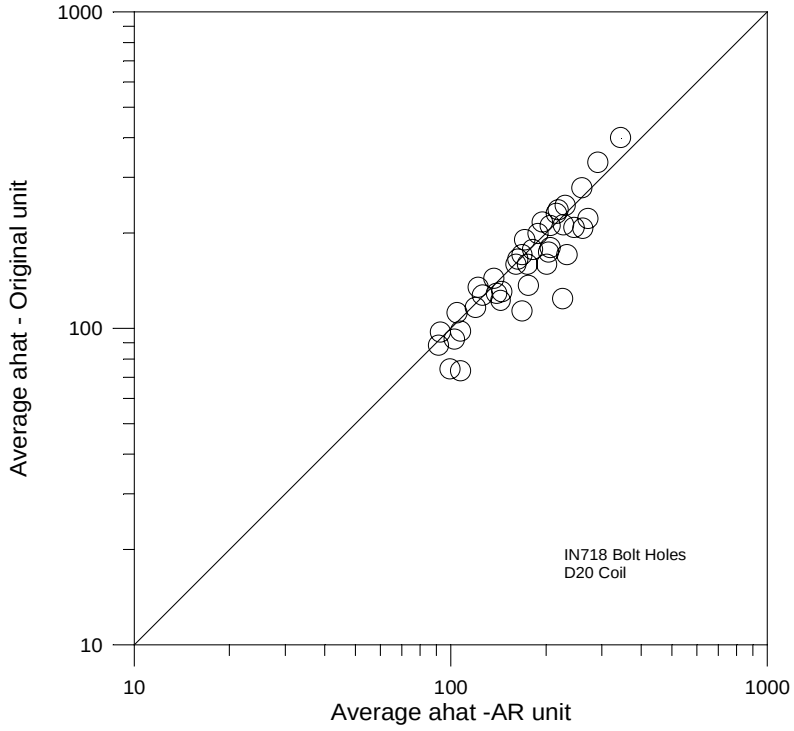


Figure 38. Comparison of $\hat{a}$ Values from Controllers – IN 718 Bolt Holes

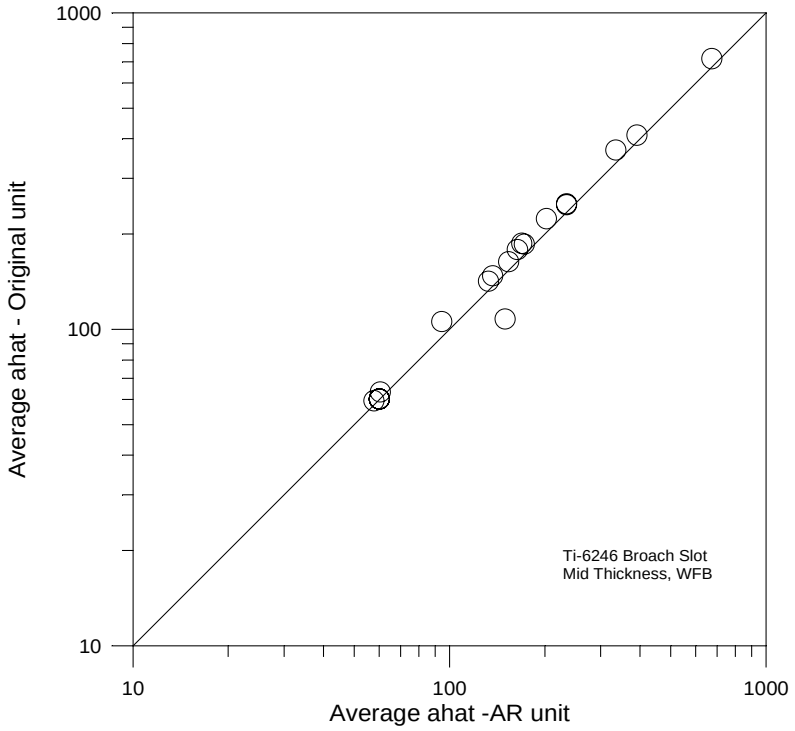


Figure 39. Comparison of $\hat{a}$ Values from Controllers – Ti-6246 Broach Slots, Mid

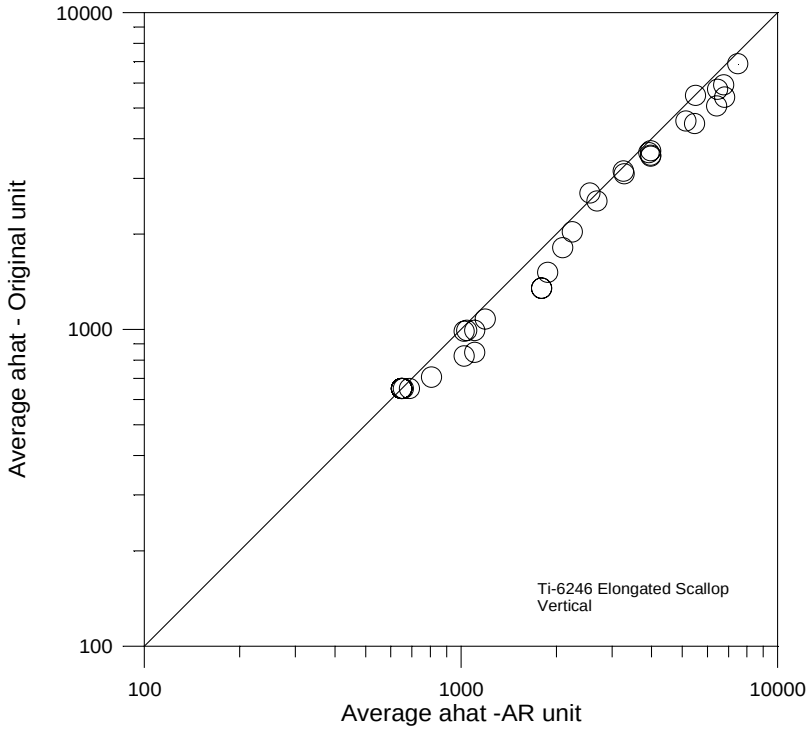


Figure 40. Comparison of $\hat{a}$ Values from Controllers – Ti-6246 Elongated Scallops

The biggest discrepancy in this set of validation tests occurred between the first and second inspections using the American Robotics unit on the Ti-6246 elongated scallops, Figure 41. No explanation could be found for the significantly larger response in the second inspection. The probe used for the second inspection was no longer available to repeat the inspection. It might be noted that the combined results from the American Robotics inspection of the elongated scallops agreed with those that had been obtained two years previously. The first inspection with the American Robotics unit agreed better with one of the inspections from the original controller than did the second inspection with the original controller.

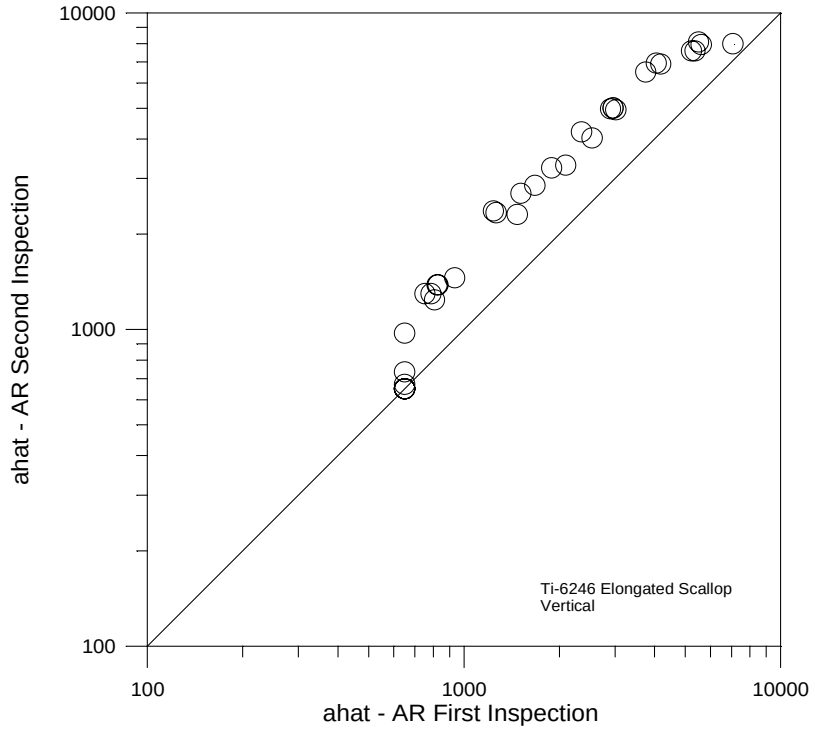


Figure 41. Comparison of $\hat{a}$ Values from American Robotics Inspections

Assuming that the cause of the difference between the two inspections of the elongated scallops was not due to the American Robotics controller and the results from this comparative study are representative, it was concluded that American Robotics controller is a drop in substitution for the original.

4.4 Integrated System Validation

A series of validation tests were conducted to demonstrate ECIS drop-in compatibility when the new eddy current instrument, station computer, and robot controller were implemented in a single integrated system [26]. The specimen sets used in these demonstration tests of the integrated system were the Waspaloy flat plate, IN718 bolt hole, Ti-6246 small bolt holes and Ti-6246 elongated scallop specimens from the F100-PW-229 program, and the Ti-6246 broach slot mid thickness from the F100-GE-220 program. All five specimen sets were inspected using the integrated ECIS system. The Waspaloy flat plate, IN718 bolt hole, Ti-6246 small bolt holes and Ti-6246 elongated scallop data were compared with the results obtained during the capability demonstration for the F100-PW-229 engine that had been conducted two years previously. The Ti-6246 broach slot specimens were inspected with the original ECIS system at the same time as inspections were made with the integrated system. The cracks in the specimen sets were inspected using two probes with the American Robotics unit and two or three probes with the original unit. The probes were not necessarily the same for the comparison within specimen sets.

The results of the comparisons are presented in Figures 42 through 51. For each specimen set, the comparison of the $\hat{a}$ values from the modified and original systems and the resulting a_{90} versus a_{thr} plots are paired. The Ti-6246 elongated scallops were analyzed for two crack size ranges and both POD(a) analyses are summarized on the same a_{90} versus a_{thr} plot. In all of the specimen sets, the differences between the original and modified systems are within the variability that has been attributed to probes, stations, and repeat inspections.

Assuming that the results obtained from the specimen sets used in this evaluation are representative of current inspections, it was concluded that the integrated system with the new eddy current instrument, station computer, and robot controller is a drop-in replacement for the preexisting ECIS system.

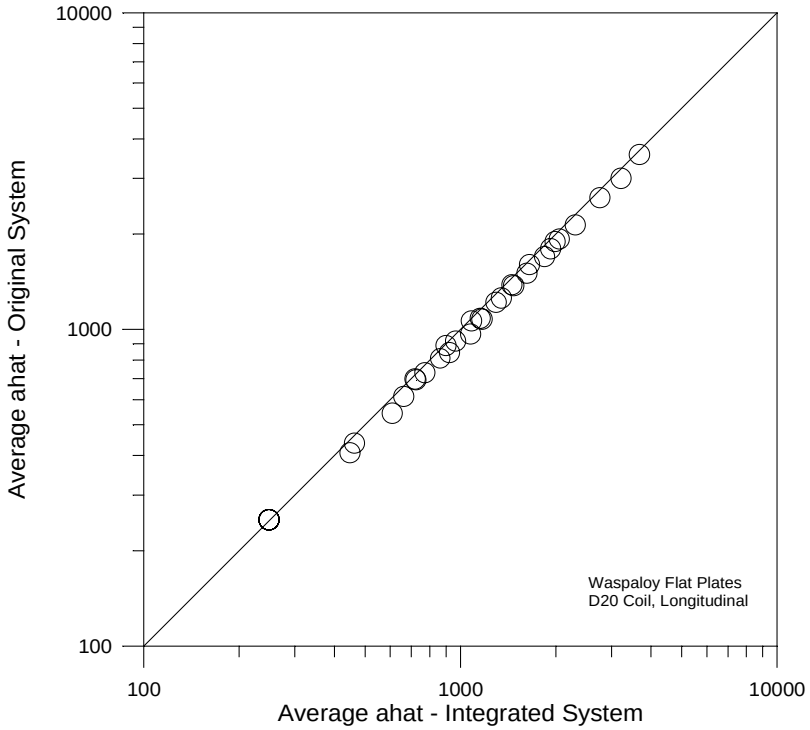


Figure 42. Comparison of Integrated System $\hat{a}$ Values – Waspaloy Flat Plates

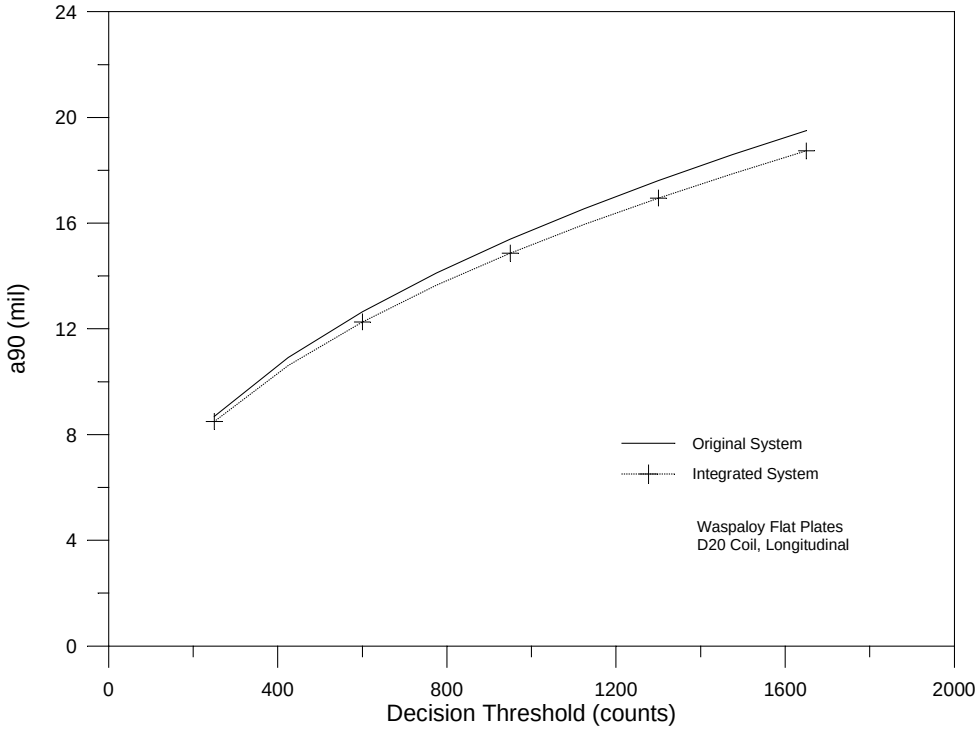


Figure 43. Integrated System a_{90} Threshold Comparisons – Waspaloy Flat Plates

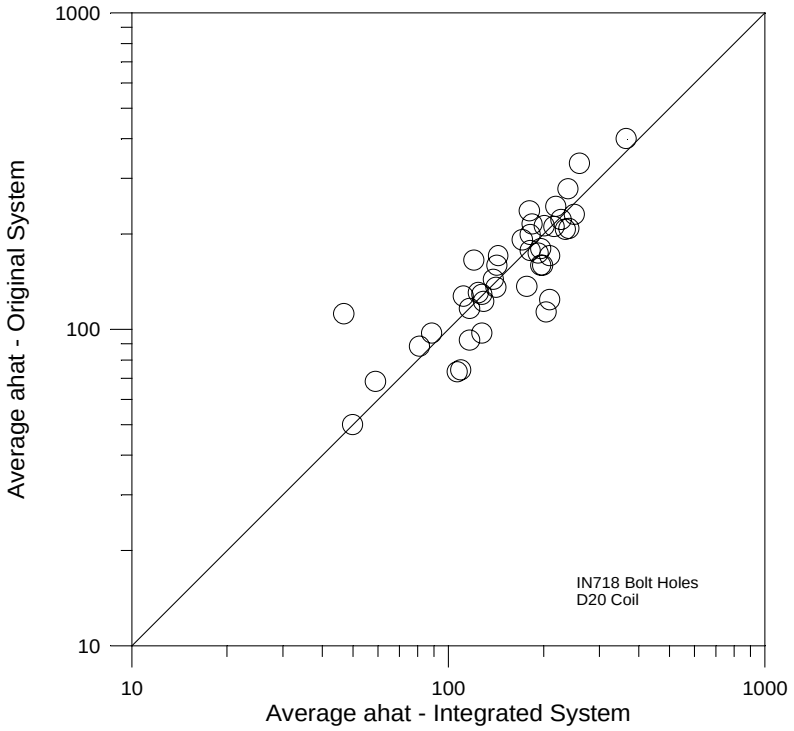


Figure 44. Comparison of Integrated System $\hat{a}$ Values – IN 718 Bolt Holes

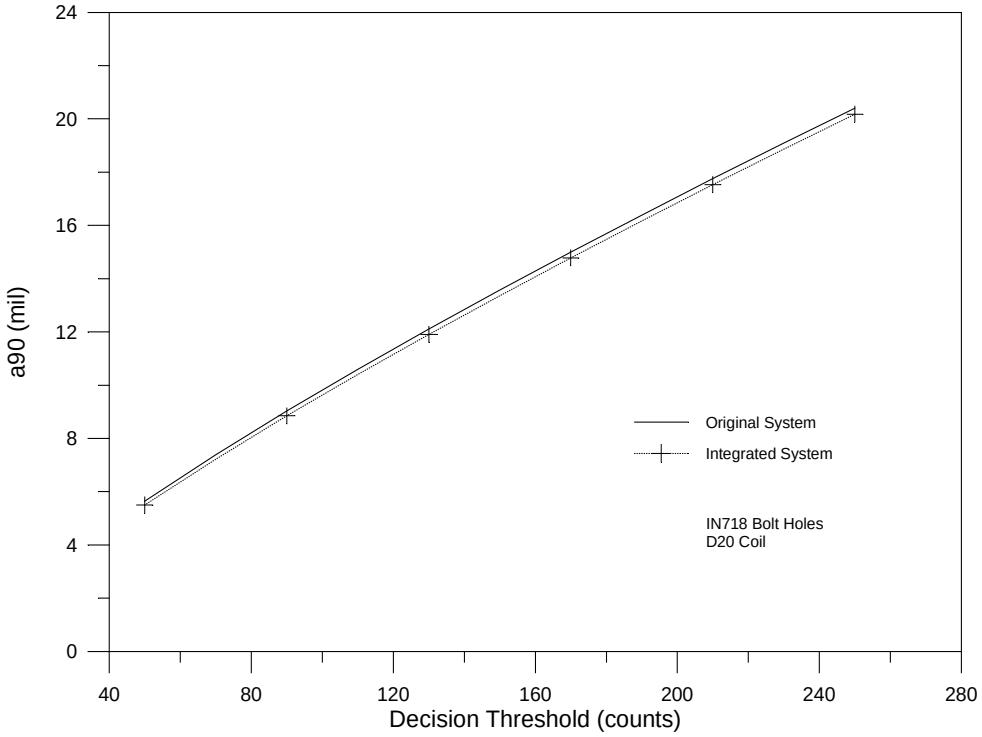


Figure 45. Integrated System a_{90} Threshold Comparisons – IN 718 Bolt Holes

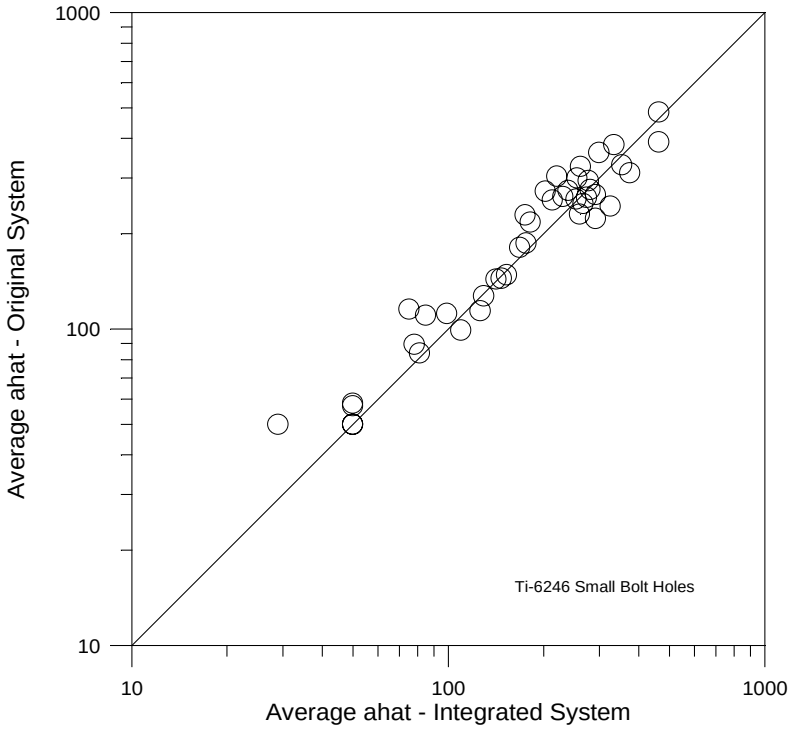


Figure 46. Comparison of Integrated System $\hat{a}$ Values – Ti-6246 Small Bolt Holes

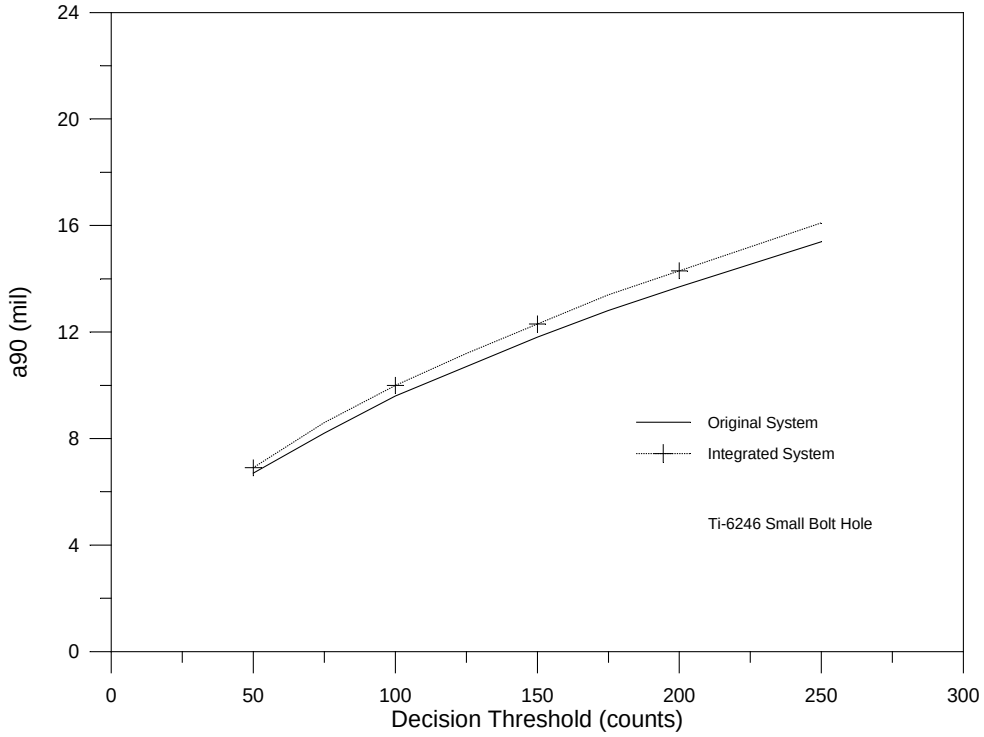


Figure 47. Integrated System a_{90} Threshold Comparisons – Ti-6246 Small Bolt Holes

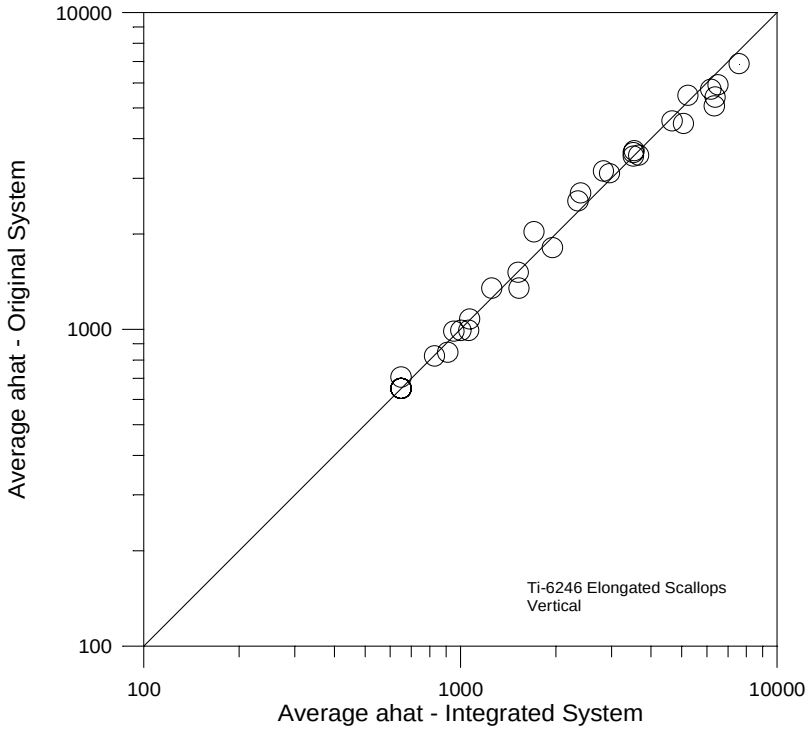


Figure 48. Comparison of Integrated System $\hat{a}$ Values – Ti-6246 Elongated Scallops

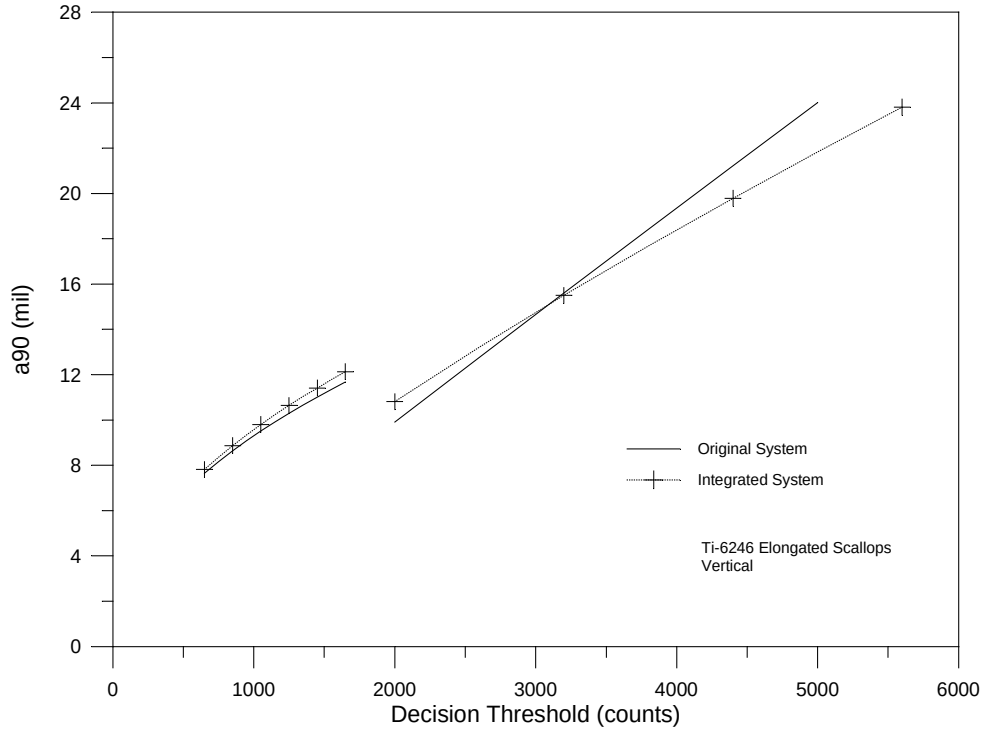


Figure 49. Integrated System a_{90} Threshold Comparisons – Ti-6246 Elongated Scallops

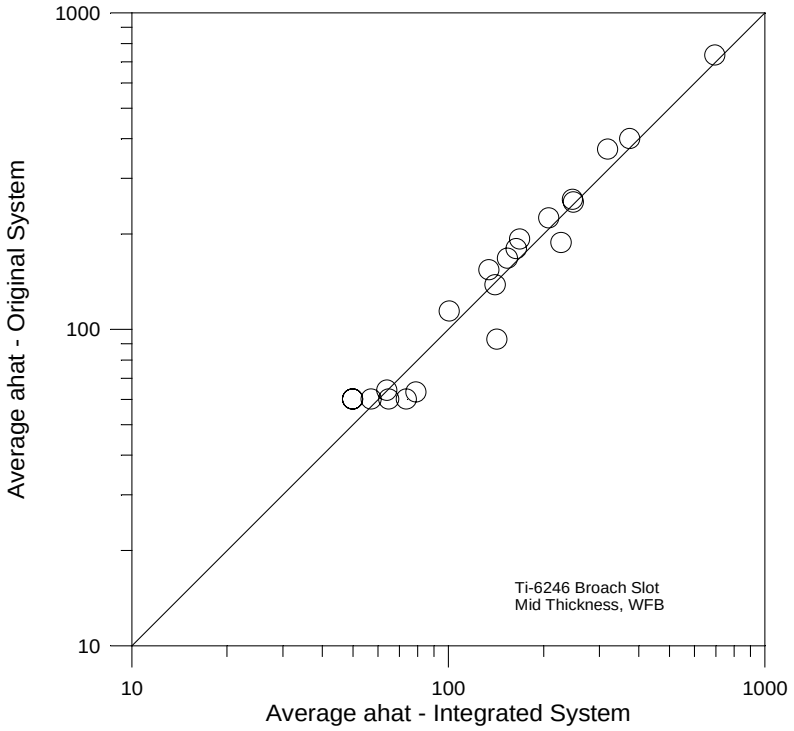


Figure 50. Comparison of Integrated System $\hat{a}$ Values – Ti-6246 Broach Slot, Mid

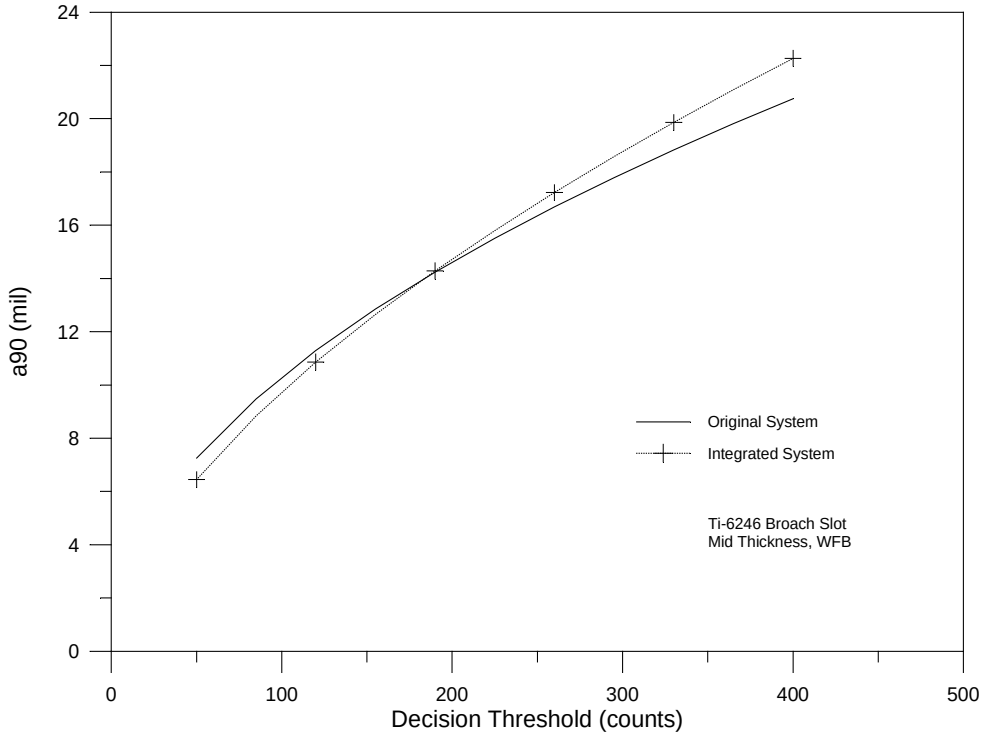


Figure 51. Integrated System a_{90} Threshold Comparisons – Ti-6246 Broach Slot, Mid

4.5 Scanner Validation

The replacement scanner [27] for the updated ECIS was validated independently of the other changes. Validation inspections were conducted using the new and original scanners on three specimen sets.

- IN718, 0.342" bolt hole specimens, D20 probe;
- Waspaloy flat plate specimens at 2 and 6 MHz, transverse orientation;
- Ti-6246 broach slot edge crack specimens at 6 MHz.

The inspection results from each specimen set are individually addressed. The validations were based on a direct comparison of the magnitudes of the system responses to the cracks in the specimens and a comparison of the resulting threshold plots for the data from the two scanners.

The IN718 bolt hole specimen set had been previously used to validate the drop-in compatibility of the integrated system modifications (subsection 4.4). Thus, the inspection results from the new scanner could be compared with the data from both the original and updated systems. Figure 52 compares average $\hat{a}$ values from the three sets of inspections and Figure 53 displays the a_{90} versus threshold plots for the three sets of inspection data. The data from the new scanner on these bolt hole specimens agree well with the other POD(a) results.

The Waspaloy flat plate specimen set was inspected in the transverse direction at 2 MHz and in the longitudinal direction at 6 MHz. The $\hat{a}$ and threshold comparisons with previous 2-MHz inspections of the specimen set are presented in Figures 54 and 55. Comparable inspection results using the original scanner were available from the validation of the eddy current instrument and a capability demonstration from the F100-PW-220 program. As seen in Figures 54 and 55, the 2 MHz inspections of the Waspaloy flat plates using the new scanner are completely compatible with those using the original scanner. However, the 6 MHz-inspection of the Waspaloy flat plate specimens with the new scanner yielded significantly different $\hat{a}$ values from those of inspection with the original scanner. Inspection results that were collected in May 1996 during the F100-PW-229 program were compared with the results from the new scanner. Figure 56 displays the difference in the responses. The $\hat{a}$ values obtained from the system with the new scanner do not decrease with crack size for the small (about 10 mil and less) cracks unlike those from the original scanner that do decrease. Note the cluster of cracks which were missed (i.e., $\hat{a}$ less than the signal threshold of 250 counts) by the system with the original scanner while the system with the new scanner yielded $\hat{a}$ values of about 1000 counts. Because the responses of the systems with the new and original scanners were so different, the POD(a) analysis was not performed. Decision thresholds would not be transferable between the systems.

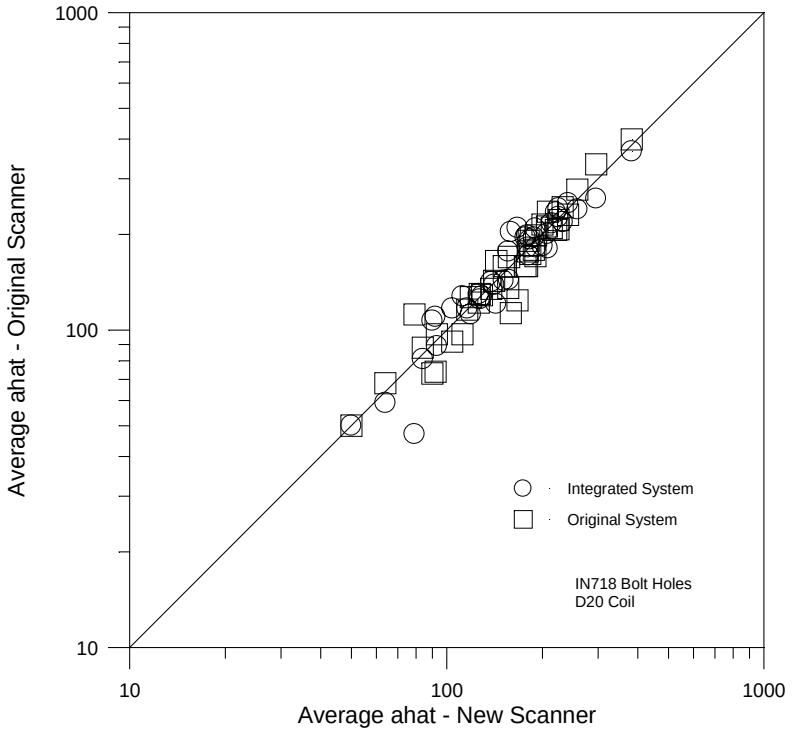


Figure 52. Comparison of Scanner $\hat{a}$ Values – IN 718 Bolt Holes

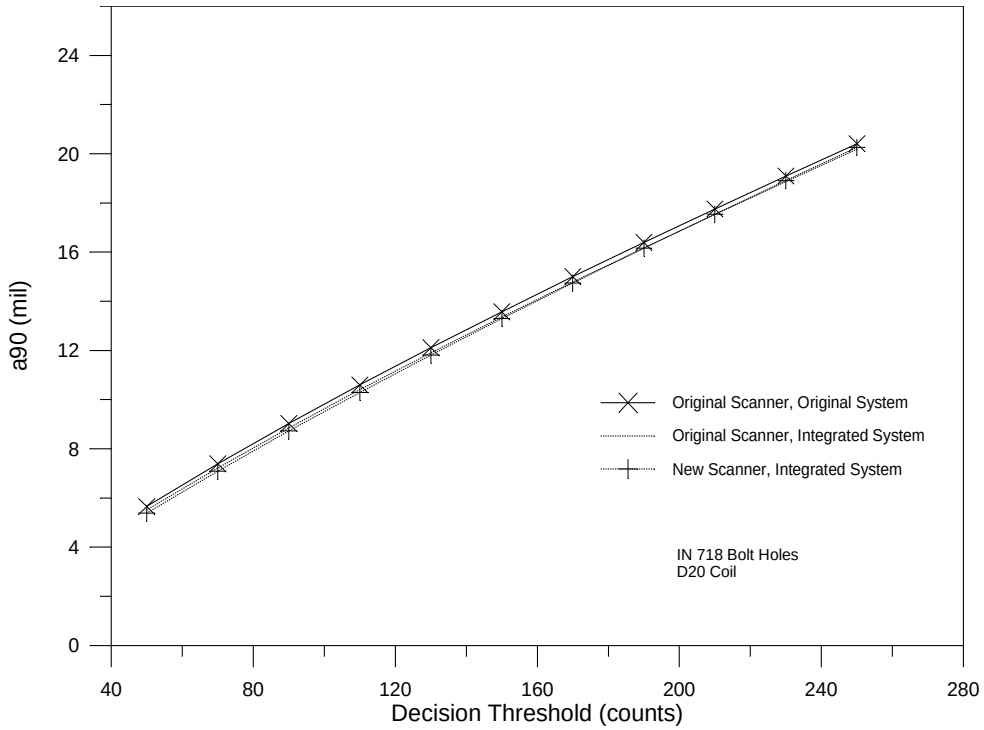


Figure 53. Scanner a_{90} Threshold Comparisons – IN 718 Bolt Holes

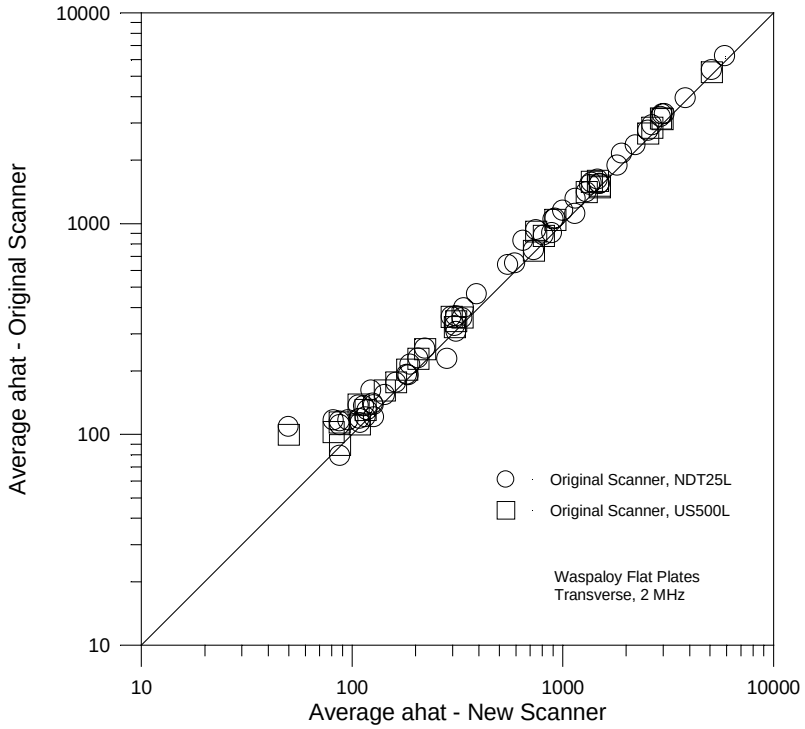


Figure 54. Comparison of Scanner $\hat{a}$ Values – Waspaloy Flat Plates, 2 MHz

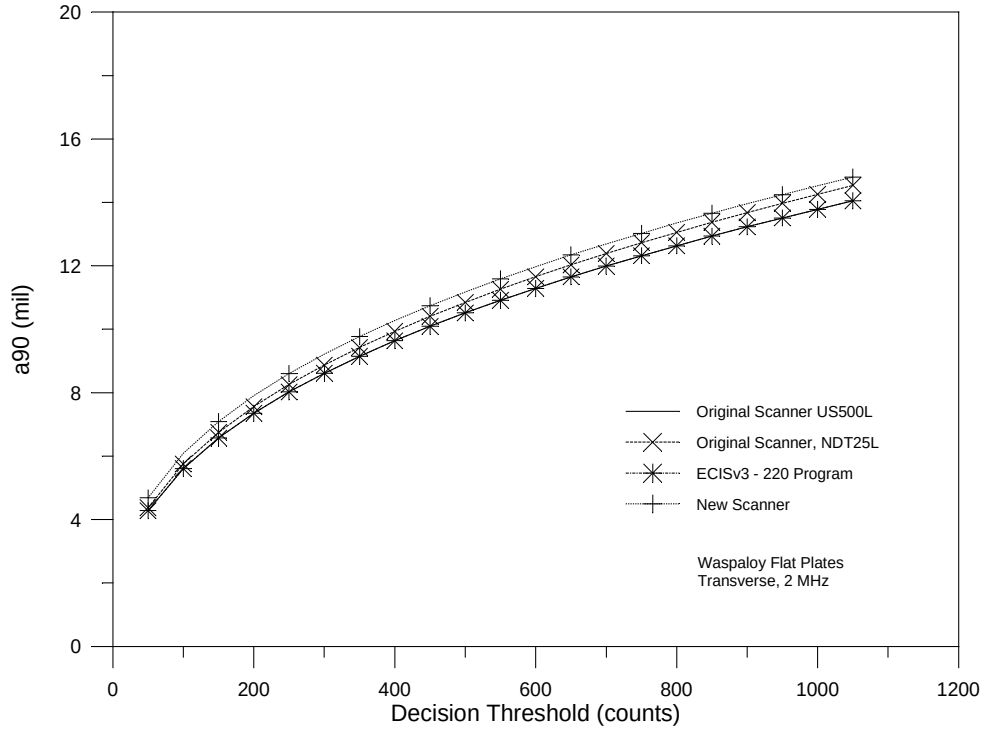


Figure 55. Scanner a_{90} Threshold Comparisons – Waspaloy Flat Plates, 2 MHz

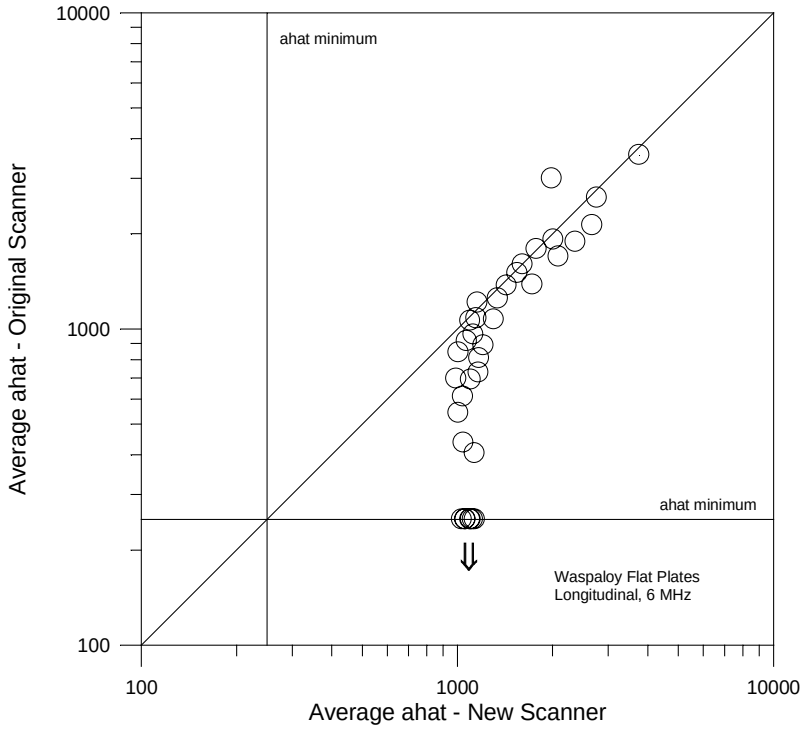


Figure 56. Comparison of Scanner $\hat{a}$ Values – Waspaloy Flat Plates, 6 MHz

The new and original scanners were used to inspect the Ti-6246 broach slot, edge crack specimens at 6 MHz. The inspections were conducted using the same station and the same two probes. Figures 57 and 58 present the comparisons of the $\hat{a}$ responses and the a_{90} versus decision threshold plots from the two inspections. The $\hat{a}$ values from the original scanner were 17 percent greater on average and there was extensive variation about this average difference. The resulting threshold plots, Figure 58, display the significantly different and nonconservative a_{90} values that would be realized if the results from the original scanner inspections were used to determine decision thresholds.

Based on the four sets of inspection data from three specimen sets, it cannot be concluded that there is a drop-in compatibility between the new and original scanners. While the 2-MHz inspections using the two scanners agreed very well, the 6-MHz inspections from the two scanners did not. In the 6-MHz inspections of the Ti-6246 broach slot edge specimens, only the scanner was different.

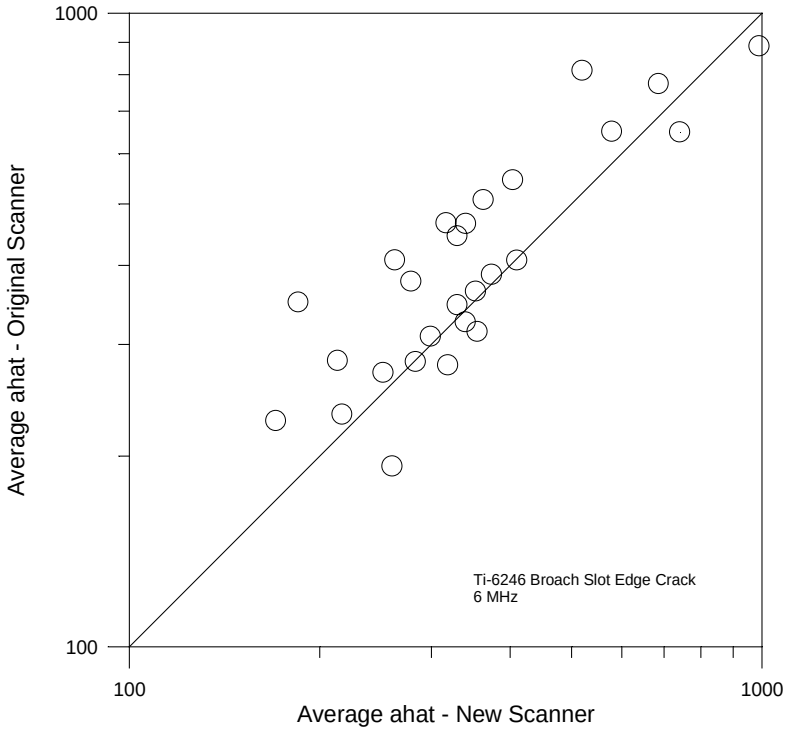


Figure 57. Comparison of Scanner $\hat{a}$ Values – Ti-6246 Broach Slot Edge Cracks

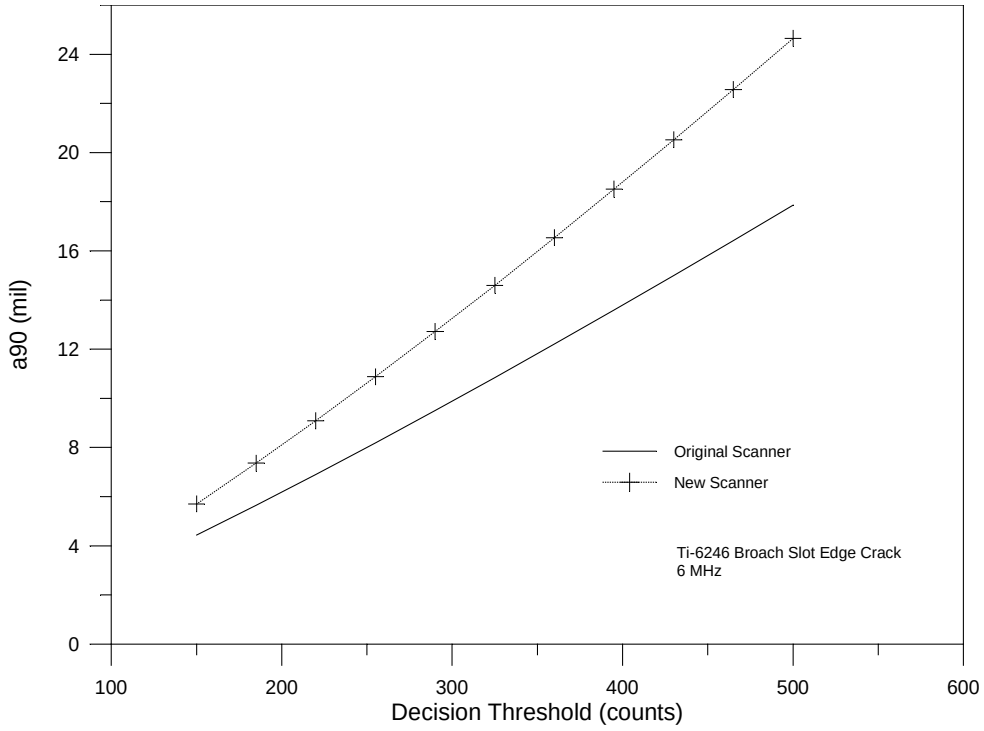


Figure 58. Scanner a_{90} Threshold Comparisons – Ti-6246 Broach Slot Edge Cracks

4.6 Calibration Method Validation

An innovative approach to calibrating the ECIS prior to inspections was developed and demonstrated as part of the upgrade program [28]. To validate this procedure, 16 cracks from a Waspaloy flat plate specimen set were inspected using four of the newly fabricated calibration blocks and the existing ECIS master calibration block for these inspections. The inspections were performed at 2 MHz and 6 MHz, and each crack was inspected with two probes and all five calibration blocks. The experimental design was completely balanced. The $\hat{a}$ data from the two frequencies were separately analyzed.

An analysis of covariance was performed to test for possible effects due to the controlled factors of the experiment. In the analysis of covariance, the effect of the $\hat{a}$ responses due to crack size is first accounted for by a regression. The effects of calibration blocks and probes on $\hat{a}$ are then evaluated by an analysis of variance. In particular, the response is modeled by the following:

$$\ln \hat{a} = B_0 + B_1 \ln (a) + C + P + (CP) + e, \qquad (40)$$

where C is the differential effect due to calibration blocks, P is the differential effect due to probes, and (CP) is the differential effect due to the interaction of calibration blocks and probes. The analysis of variance first determines if there is any effect due to the calibration blocks. If so, follow-up analyses would isolate the differing blocks. The possible effects due to calibration blocks and to the interaction between calibration blocks and probes were not significant in both the 2 and 6 MHz data sets. The only statistically significant effect was that due to the probes in the 2 MHz inspections. Figures 59 through 62 present the $\hat{a}$ values from the new calibration blocks plotted against values obtained from the master block for the four combinations of frequency and probe. The slightly, but statistically significant, greater ahat values from the 2MHz, Probe #2 data can be seen in Figure 60. To further illustrate this probe difference, Figure 63 compares the $\hat{a}$ values from the two probes using only the ECIS master calibration blocks. This multiplicative shift is due to variability in response between probes and recalibrations.

Although 16 cracks are not considered to be a sufficiently large number for POD(a) analysis, a_{90} values were estimated for each of the combinations of frequency and calibration block. The a_{90} values from the new calibration blocks were within 1.5 mil of those from the ECIS master blocks across the entire crack size range in the specimens. In fact, the biggest discrepancy resulted from the probe difference on the 2 MHz inspections using ECIS master calibration blocks.

Assuming that the inspections of the Waspaloy flat plates is representative of general ECIS inspections, it is concluded that inspection results using the new method of calibration would agree with using the current ECIS master calibration blocks.

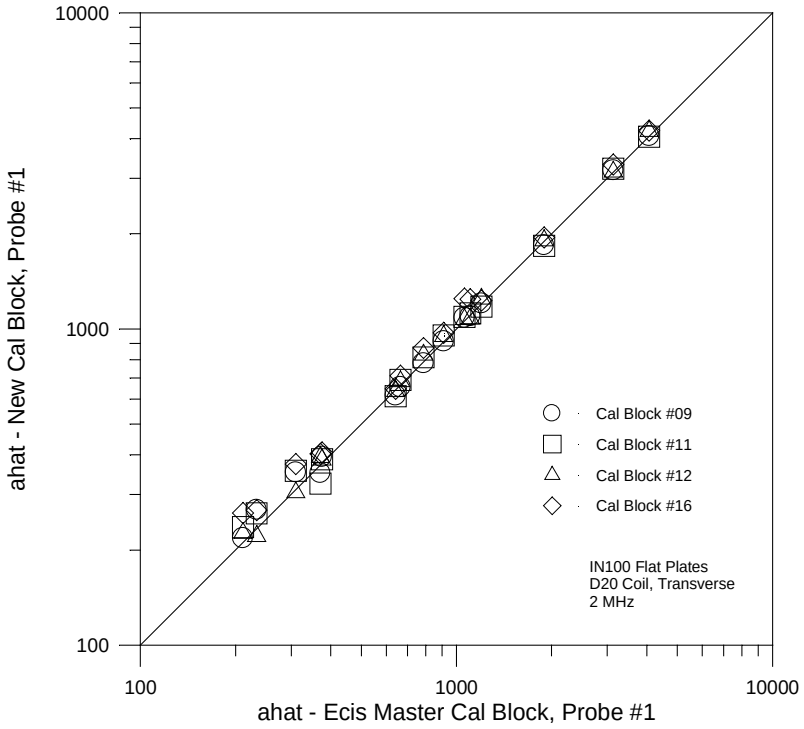


Figure 59. Calibration Comparison with ECIS Master Block – 2 MHz, Probe #1

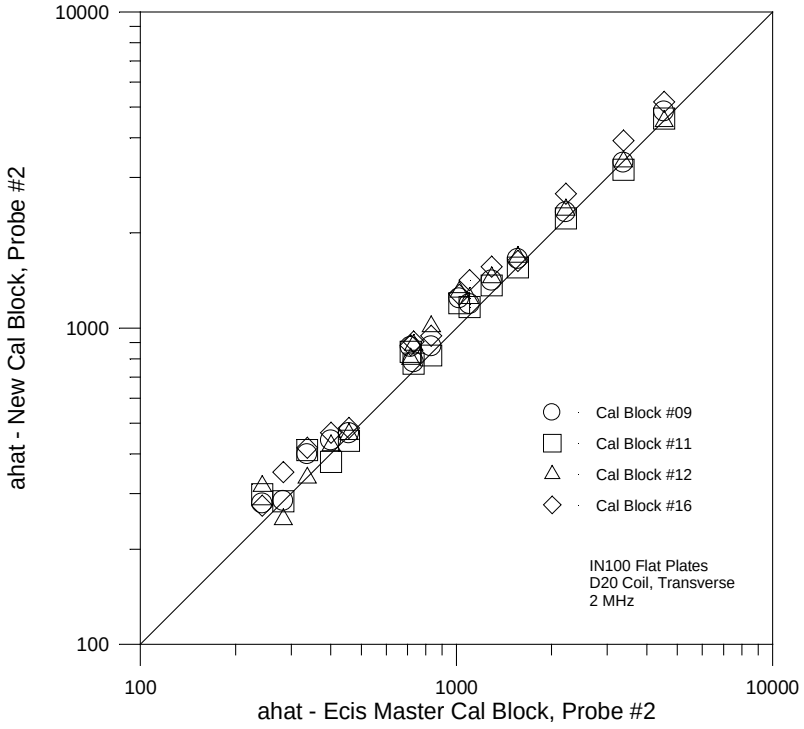


Figure 60. Calibration Comparison with ECIS Master Block – 2 MHz, Probe #2

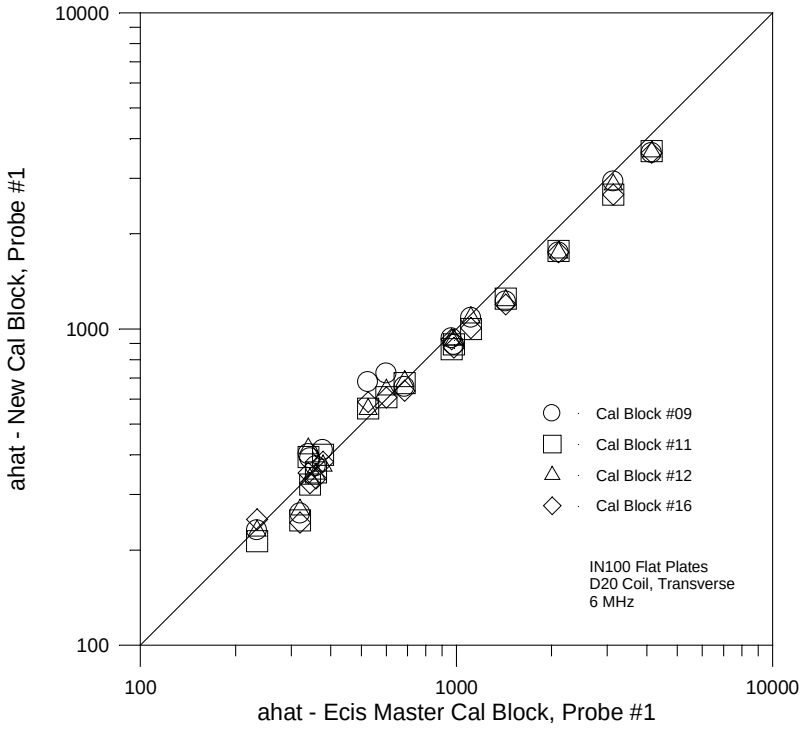


Figure 61. Calibration Comparison with ECIS Master Block – 6 MHz, Probe #1

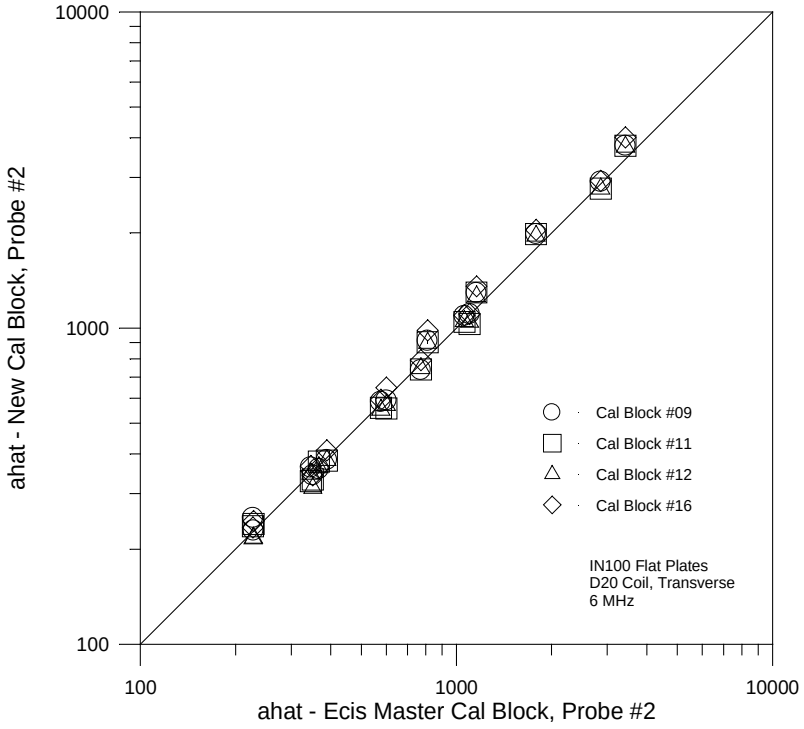


Figure 62. Calibration Comparison with ECIS Master Block – 6 MHz, Probe #2

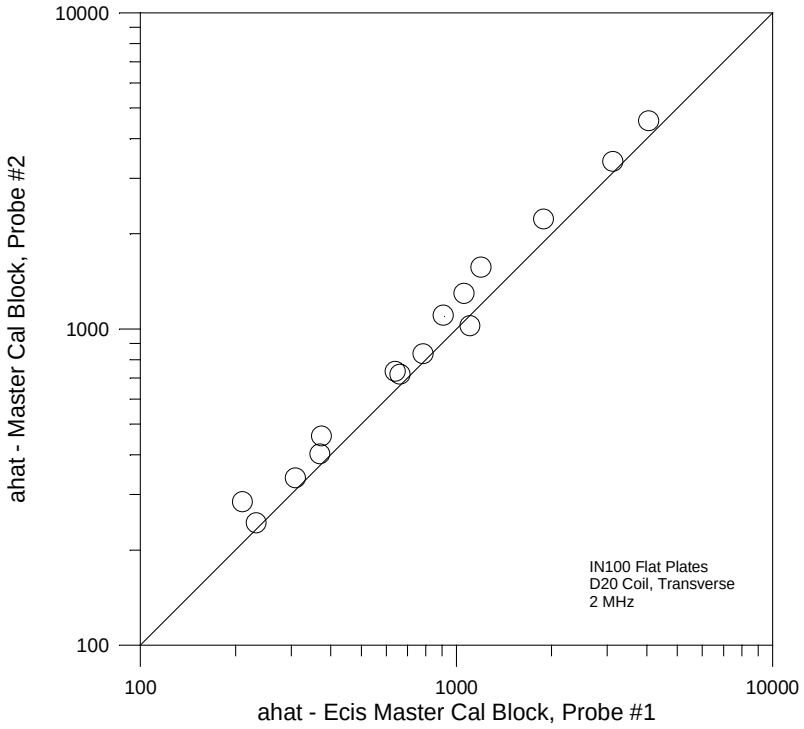


Figure 63. ECIS Master Block Probe Variability – 2 MHz

Section 5
Summary and Conclusions

The ECIS is a highly automated nondestructive inspection system that is an essential component of the U.S. Air Force RFC/ENSIP. The ECIS was initially developed for the U.S. Air Force in the 1980's. As part of a program to upgrade the hardware and software of the ECIS, the computer code for evaluating the capability of NDE systems in terms of POD was updated and the ECIS modifications were evaluated for drop in compatibility with the existing system. This report described the changes to the POD computer program and presented the results of the POD evaluation of the drop-in compatibility of the system modifications.

The computer programs for performing the statistical calculations required by a POD analysis based on a cumulative lognormal or normal model was completely rewritten. The new program, designated POD Version 3, uses Microsoft Excel as the interface for input and output to C^{++} programs that find the maximum likelihood estimates of the POD(a) parameters. In addition to identifying information and parameter estimates, the standard output includes statistical tests of the validity of the assumptions required for the model. At the discretion of the analyst, the output can also include a POD(a) plot with confidence bound and plots of the fit for both the $\hat{a}$ versus a and pass/ail analyses. For an $\hat{a}$ versus a analysis, plots of $\hat{a}$ versus a, detection threshold versus a_{90}, and $\hat{a}$ residuals versus a can also be generated. POD Version 3 also permits transformations of crack size or $\hat{a}$. The analyses can be performed in terms of no, logarithmic, inverse or other user-defined transformation. The same transformations can be applied to the inspection response, $\hat{a}$, when available. The logarithmic is the default transformation for both crack size and $\hat{a}$.

Three ECIS upgrades were completed during the program. These included the eddy current instrument, the ECIS station computer, and the robotic controller. In addition, a novel approach to calibration was devised and demonstrated. All of these modifications were evaluated in terms of potential effects on POD to demonstrate that previously set detection thresholds would still be valid after a change to the new equipment. The eddy current instrument, station computer, and robotic controller were evaluated both individually and in a single system containing all three updates. The drop-in compatibility was validated by a direct comparison of $\hat{a}$ values from identical inspections and/or by comparing decision threshold versus a_{90} plots that resulted from inspecting common specimen sets. In all of these cases, it was concluded that variability in system response due to the updated components was within the expected variability that is attributable to other non-controllable sources. The system upgrades are drop-in compatible, and the new calibration method yields $\hat{a}$ responses that are completely compatible with those of the original system.

A replacement scanner was developed and evaluated late in the program. Data collected on IN718 bolt holes and at 2 MHz on Waspaloy flat plates using the new and original

scanners agreed well. However, data collected at 6 MHz on the Waspaloy flat plates and on titanium broach slot edge crack specimens displayed significant differences between the two scanners. The new scanner as evaluated did not provide signal responses that were compatible with the old scanner.

Section 6
References

1. Harris, J.A., Jr., Sims, D.L., and Annis, C.G., Jr., "Concept Definition: Retirement for Cause of F100 Rotor Components," AFWAL-TR-80-4118, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, September 1980.

2. Mil-HDBK-1783 (USAF), "Engine Structural Integrity Program, (ENSIP)," 30 November 1986.

3. Stubbs, D.A. and Hoppe, W.C., "RFC Automated Inspection Overview," Review of Progress in Quantitative Nondestructive Evaluation, Vol. 5A. Edited by D.O. Thompson and D.E. Chimenti, Plenum Press, 1985, pp. 901-910.

4. Mil-HDBK-1823, "Non-Destructive Evaluation System Reliability Assessment," 30 April 1999.

5. Berens, A.P., "NDE Reliability Data Analysis," ASM Metals Handbook, Volume 17, 9th Edition: Nondestructive Evaluation and Quality Control, ASM International, Materials Park, Ohio, 1988, pp. 689-701.

6. AFGS-87221A, Air Force Guide Specification, Aircraft Structures, General Specification for, 8 June 1990.

7. Petrin, C., Annis, C., and Vukelich, S.I., "A Recommended Methodology for Quantifying NDE/NDI Based on Aircraft Engine Experience," AGARD-LS-190, Advisory Group for Aerospace Research and Development, NATO, Neuilly Sur Seine, France, April 1993.

8. Spencer, F.W. and Schurman, D.L., "Reliability Assessment at Airline Inspection Facilities, Volume III: Results of an Eddy Current Inspection Reliability Experiment," DOT/FAA/CT-92/12, III, FAA Technical Center, Atlantic City, NJ, March 1994.

9. Lewis, W.H., Sproat, W.H., Dodd, B.D., and Hamilton, J.M., "Reliability of Nondestructive Inspections – Final Report," SA-ALC/MME 76-6-38-1, San Antonio Air Logistics Center, Kelly Air Force Base, Texas, December 1978.

10. Sproat, W.H., Hamilton, J.M., and Hovey, P.W., "Engineering Services in Support of NDI Operations," SA-ALC/MMEI/87-01, NDI Program Office, San Antonio Air Logistics Center, Kelly AFB, Texas, October, 1988.

11. Hovey, P.W., Sproat, W.H., and Schattle, P., "The Test Plan for the Next Air Force NDI Capability and Reliability Assessment Program," Review of Progress in Quantitative Nondestructive Evaluation, Vol. 8B. Edited by D. O. Thompson and D. E. Chimenti, Plenum Press, 1989, pp. 2213-2220.

12. Davis, M.K., "Proficiency Evaluation of NDE Personnel Utilizing the Ultrasonic Methodology," Review of Progress in Quantitative Nondestructive Evaluation, Vol. 7B. Edited by D. O. Thompson and D. E. Chimenti, Plenum Press, 1988, pp. 1777-1789.

13. Berens, A.P. and Hovey, P.W., "Evaluation of NDE Reliability Characterization," AFWAL-TR-81-4160, Volume 1, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, December 1981.

14. Berens, A.P. and Hovey, P.W., "Flaw Detection Reliability Criteria, Volume I – Methods and Results," AFWAL-TR-84-4022, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, April, 1984.

15. Berens, A.P. and Hovey, P.W., "The Sample Size and Flaw Size Effects in NDI Reliability Experiments," Review of Progress in Quantitative Nondestructive Evaluation, Vol. 4B. Edited by D.O. Thompson and D.E. Chimenti, Plenum Press, 1985, pp. 1327-1334.

16. Box, G.E.P. and Cox, D.R., "An Analysis of Transformations," Journal of the Royal Statistical Society, Series B, 26, pp. 211-252, 1964.

17. Neter, J., Wasserman, W., and Kutner, M.H., Applied Linear Regression Models, Richard D. Irwin, Inc., Homewood Illinois, pp. 132-140, 1983.

18. Hosmer, D.W. and Lemeshow, S., Applied Logistic Regression, John Wiley & Sons, New York, 1989.

19. Lawless, J.F., Statistical Models and Methods for Lifetime Data, John Wiley & Sons, New York, pp. 225-226, 1982.

20. Draper, N.R., and Smith, H., Applied Regression Analysis, 2nd Edition, John Wiley & Sons, New York, pp. 33-40, 1981.

21. Ostle, B., Statistics in Research, The Iowa State College Press, Ames, Iowa, p. 242, 1954.

22. D'Agostino, R.B. and Stephens, M.A., Goodness-of-Fit Techniques, Marcel Dekker, Inc, New Marcel Dekker, Inc, New York, pp. 372-373, 1986.

23. Leethy, M., "Eddy Current Instrument Final Report," Task 1, Contract Number F33615-95-C-5236, Veridian Engineering, Dayton, Ohio, 2000.

24. Olding, R., “Final Report, New ‘Open’ Inspection Station Computer,” Task 5, Contract Number F33615-95-C-5236, Veridian Engineering, Dayton, Ohio, 2000.

25. Collins, P, “Retirement for Cause/Engine Structural Integrity Program, Advanced Capability Motion Controller,” American Robotics, 2000.

26. Braun, T., “PRDA Component Integration Acceptance Test Report,” Contract Number F33615-95-C-5236, Veridian, Engineering Dayton, Ohio, 2000.

27. Leethy, M., “Scanner Replacement Final Report,” VE-PRDA-103, Veridian Engineering, Dayton, Ohio, 2000.

28. Stubbs, D. A., Martin, R.W., Schell, N.D., and Petricola, D.L.,“Retirement for Cause (RFC) Eddy Current Inspection System Calibration Improvement,” UDR-TR-2000-00057, University of Dayton Research Institute, Dayton, Ohio, February 2000.