Three Decades of NDI Reliability Assessment

Report No. Karta-3510-99-01
Submitted to AF NDI Office
Contract F41608-99-C-0404

May 2000

Ripudaman Singh

5555 Northwest Parkway, San Antonio TX 78249
Http://www.karta.com/POD

FOREWORD

This report reviews three decades (1970-1999) of engineering and research efforts to quantify capability and reliability of non-destructive inspections in aerospace and other industries. Karta Technologies, Inc. San Antonio, TX performed the review under Air Force NDI Office contract F41608-99-C-0404, "Development of Probability of Detection Protocol Program Plan and Pilot Program (Personnel Assessment) Identification/ Initiation." Donald Locke was the Project Manager from Karta and Mike Paulk is the Air Force NDI Program Manager. Robert R Lewis is the AF Project Officer. The task was performed during July-October 1999 and covers nearly 150 reports and manuscripts from over 100 authors. Experts on the subject reviewed the report during Dec 1999-March 2000.

The Air Force provided an initial set of 15 reports on major U.S. programs. Karta acquired the remaining reports from NTIAC, DTIC, SwRI, and its own library. All reports related to the subject under investigation and accessible to Karta within reasonable and prudent means were covered. However, Karta does not claim to have reviewed every report that exists.

This review document is organized rather unconventionally and the format is discussed in Section 1.4.

Acknowledgements: Karta Tech. acknowledges the general guidance from Mike Paulk (AF NDIO), Bob Lewis (AF NDIO), Ward Rummel (D&W Enterprise), Floyd Spencer (FAA-AANC, Sandia National Labs), and Library Personel at DTIC, NTIAC, and SwRI were extremely cooperative. Carol Liebowitz (AF), Ward Rummel, Bill Sprout (Ex. Lockheed), Floyd Spencer, Gopi Katragadda (Karta), and Mahlon Long (Karta) provided critical review of the document. Shirley Heller helped with Technical editing. Dan Palumbo (Karta) provided support with search and acquisation of reports reviewed. The administrative support from Elaine Lewis (Karta) is appreciated.

Karta Technologies, Inc.

San Antonio, TX 78249

May 2000

SUMMARY

Reliability of a non-destructive inspection (NDI) procedure has been defined as a quantitative measure of the efficiency of that procedure in finding flaws of specific type and size (Metals Handbook). Damage tolerant design requires knowledge of the reliability of the inspection technique used to detect flaws or damage. Over the last three decades many NDI reliability assessment programs were conducted, and various quantitative measurements were employed to express the NDI system capabilities.

In the initial stages of NDI assessment, the metric was the 90/95 crack length, which is defined as the minimum crack length for which there is a probability of detection of 90% with 95% confidence. The reliability presentations then moved to full probability of detection (POD) curves typically at 95% and 50% confidence levels. Subsequently other metrics and representations have been used to account for the probability of false calls (e.g., Relative Operating Characteristic curves and Coefficient of Contingency). The experimental efforts range from a simple set of specimen tests at one facility to a truckload of realistic structures traveling around the country for months together. Statistical tools have evolved from basic calculations which required a large number of data points (e.g., no. of finds/no. of flaws) to those with mathematical rigor but requiring fewer data points (e.g., log-odds).

Most of the studies identified controllable as well as non-controllable factors that clearly effect NDI reliability. These include the material, process, equipment, procedure, and more importantly the human. Many investigations attributed the variations observed in detection capability to the human element. Human factors have become a subject of great interest and deserve an in-depth investigation.

Over the past 10 years, efforts have been made to develop mathematical and numerical models for prediction of POD curves to reduce the experimental validation requirements. Although their success was limited, they do show tremendous potential. Evolution of computational procedures offer advances not only in data manipulation, but also in the form of actual inspection simulation. In the future, the role of computers will make reliability assessment programs quicker, cheaper, more accurate, and more dependable.

Reliability assessment of an aircraft or engine inspection involves many continuously changing variables, including the unpredictable human, that it will continue to remain a challenge. This report is an attempt to consolidate the vast amount of valuable information available from past efforts of many engineers and organizations who have made remarkable contributions to the field of reliability assessment.

TABLE OF CONTENTS

ACRONYMS

AANC	Aging Aircraft NDI Center
AF, AFB	Air Force, Air Force Base
ALC	Air Logistics Center
CC	Coefficient of Contingency
CL	Confidence Limits
DTIC	Defense Technical Information Center
ECIRE	Eddy Current Inspection Reliability Experiment
EDM	Electronic Discharge Machine
ENSIP	Engine Structural Integrity Program
EPRI	Electric Power Research Institute
ET	Eddy Current Testing
FAA	Federal Aviation Administration
IGSCC	Intergranular Stress Corrosion Cracking
MOI	Magneto-Optic eddy current Imaging
MPI	Magnetic Particle Inspection
MT	Magnetic Particle Testing
NBC	Nuclear, Biological and Chemical
NDE, NDI, NDT	Nondestructive Evaluation, Inspection, Testing
NRC	Nuclear Regulatory Council
NTIAC	NDT Information Analysis Center
PML	Percentage Material Loss
PNL	Pacific Northwest Lab
POD	Probability of Detection
POFA	Probability of False Alarm
POI	Probability of Indication
PT	Penetrant Testing
RFC	Retirement for Cause
ROC	Relative Operating Characteristics
RT	Radiographic Testing
SAIC	Science Application International Corporation
SwRI	Southwest Research Institute
T.O.	Technical Order
UDRI	University of Dayton Research Institute
UT	Ultrasonic Testing

1. INTRODUCTION

1.1 Preamble

The need for higher safety has shifted the design philosophy for new structures from safe-life or fail-safe to Damage Tolerance (i.e., *safety by inspection).* A structure is termed as damage tolerant if it has a reasonable damage growth life such that the damage can be detected during one of the scheduled inspections before it can precipitate a failure. Figure 1.1 depicts a graphical representation of the damage tolerance concept. In the damage tolerance philosophy, the inspection intervals are governed by the ability to detect smaller cracks and safely tolerate large cracks. Also, life-extension programs of aging structures are increasingly dependant on damage detection and repair, irrespective of the original safe design criterion. The economic and safety requirements rely substantially on the capability to detect and characterize cracks and other defects in service. Quantitative assessment of field NDI capability and reliability is thus a major component of continued safe fleet operations.

Over the years, two approaches to quantification of NDI[1] have evolved: (1) demonstration of capabilities at a fixed (assumed) flaw size and (2) characterization of NDI procedures by a Probability of Detection curve (POD). A POD curve is generally produced by applying an NDI procedure to a large number of cracks in the production environment, correlating the results of the inspection with each crack size, analyzing the data, fitting the results to a model, and plotting the results as a function of flaw size [Easter 98]. Figure 1.2 qualitatively shows a POD curve.

The POD method evolved in the late 1960's and early 1970's to meet the design and production needs of the NASA Space Shuttle and the Air Force B-1 Bomber program. These programs constituted the first major use of fatigue and fracture mechanics as a basis for design and life cycle maintenance. The NASA approach provided more information using methodologies that were consistent with the generation of materials properties data. A joint meeting between NASA, Air Force Material Laboratories personnel, and the contractors resulted in agreement to use the NASA methodologies for reporting NDI reliability. The POD reporting method is now considered to be a standard method for quantification of NDI process performance. It can be used to include: NDT procedure validation, NDT personnel proficiency demonstration and qualification, comparative analysis of NDT processing materials, equipment and procedures, appropriate NDI method selection, automated NDT systems qualification, design

[1] In this document, the terms NDI (Nondestructive Inspection), NDE (Nondestructive Evaluation), and NDT (Non destructive Testing) mean the same process: nondestructive sampling a material or a part using a method to check the soundness of that material or part without impairing or destroying the serviceability. Different programs in the past have used the terms NDI, NDE, NDT appropriately to represent their activity. This document attempts to maintain the original rpresentation.

acceptance requirement establishment, and inspection interval estimation. Engineering reference to POD data generated in the aerospace industry is available in the form of an *NDE Capabilities Data Book* [Matzkanin 97]. The practical application of an NDI procedure is limited by the ability of the method to discriminate between signal and noise [Rummel 83]. If the discrimination threshold is set too high, flaw will be missed and the POD reduce. If the discrimination threshold is set too low, noise will be reported as a flaw signal and Probability of False Alarm (POFA) will increase. The POFA is a significant economic consideration in characterizing the performance of an NDI procedure. An interrelationship between POD and POFA can be described in the form of a Relative Operating Characteristics (ROC) curve. This form of analysis was first introduced during World War II as a method of assessing the discrimination capabilities of radar operators in response to scope signal indications [Tanner 54]. Figure 1.3 qualitatively shows an ROC curve as applied to NDI data.

The inspection capability assessment involves observing the condition of facilities, the materials and equipment, operating practices, quality and process control of the NDI procedure and overall equipment and operator competence in execution of NDI process. Reliability determinations measure flaw detection probabilities derived by NDI and the ability of man-equipment to discriminate between signal and noise to recognize and discern flaw characteristics [Hovey 89].

1.2 Current Objectives

USAF Program Objectives

NDI performed by field-level personnel is an important activity in sustaining military readiness and effectiveness of operational aircraft. To help ensure that NDI activities continue to meet Air Force objectives, the AF NDI Office undertook development of a Generic Protocol Program Plan for Air Force wide NDI assessment. The protocol will provide a common basis of assessment by generating accurate assessments for individuals, as well as larger organizational units (e.g., entire labs or commands). A common basis of assessment for each element of the NDI system allows an accurate evaluation of the impact on mission readiness and consistent comparison across NDI elements; it also provides a basis for determining and evaluating proposed improvements.

The Air Force objectives for the overall program, of which this project is a part, are to:

1. Develop the hardware/software technology, procedures, instructions, and protocol for AF NDI assessment on a regular basis.
2. Implement the protocol to assess the NDI performance at various AF NDI labs and depots as a measure of routine NDI effectiveness, and new technologies, equipment, and applications.
3. Compare the measured NDI performance against requirements and acceptable standards.

4. Identify NDI system deficiencies that cause performance short fall.
5. Establish need for improvement and make necessary recommendations to improve NDI effectiveness.
6. Continue program implementation and program improvements keeping pace with developing technology and evolving operational requirements.

Project Objectives

The objective of the Karta project is to develop the Protocol Program Plan for future demonstration and evaluation by the USAF NDI Office. The project spans one year starting in June 1999 and consists of the following four major tasks:

Task 1: Conduct a study of earlier efforts to quantify NDI performance.

Task 2: Conduct a survey of human factors with active-duty NDI inspectors.

Task 3: Develop requirements for NDI performance assessment.

Task 4: Develop the Protocol for conducting NDI performance assessment.

The present report addresses Task 1: Review of past NDI reliability assessment efforts.

Task Objectives

The objectives of Task 1 (Review of past efforts) are as follows:

1. Gather available literature from earlier efforts and understand the past activities.
2. Compile the information and identify those successful models that can be reapplied.
3. Identify those models that show value but limited success.

1.3 Reliability Assessment Program Elements

In general, NDE comprises the application of a stimulus to a structure and the interpretation of the response to the stimulus. Many factors influence whether or not the inspection will produce a response that will result in a correct decision as to the absence or presence of a flaw. Repeated inspections of a specific flaw can produce different magnitudes of stimulus responses because of minute variations in equipment setup and calibration and slight differences in flaw characteristics. These variations are inherent in the end to end NDI process. Different flaws of the same size can produce different response magnitudes because of differences in material properties, flaw geometry, and flaw orientation. The interpretation of the response can be influenced by the capability of the interpreter (man or machine), the mental acuity of the inspector as influenced by fatigue or emotional outlook, and the ease of access and the environment at the inspection site. All these factors contribute to the inspection uncertainty and lead to probabilistic characterization of the inspection capability [Berens 88].

The objective of quantifying NDI capabilities is generally to relate the output of the NDI process/procedure to a desired or an undesired characteristic of the test object. Inspections are performed to "detect cracks" as a characteristic of primary importance. Although other characteristics are measured and assessed, crack detection usually is the primary focus of data presentation. Quantification of detection as a function of crack size is the output of most NDI capability characterizations. A single-valued parameter that characterizes a procedure is the crack size at which POD reaches the 90% level. The single-valued parameter that has been often quoted in validation requirements is that crack size at which the POD is 90% with a 95% confidence level, commonly referred as 90/95 crack size.

The design of an NDI reliability assessment program needs special care in order to produce "good data" and provide confidence in the results obtained. Traditionally, it has involved: (1) selecting an adequate number of representative structures with statistically significant number of known defects, (2) subjecting the damaged structural specimens to routine NDI at a predetermined AF depot and field installations with a representative sample of NDI technicians and inspection conditions, and (3) analyzing the results of NDI in terms of flaw detection probabilities as a function of flaw size.

Computational models now can predict NDI reliability and also simulate the NDI response signal. We are likely to see more simulation-based NDI reliability and capability assessment programs in the future.

1.4 Report Format

Since the early 70s, various organizations performed almost a dozen significant programs to assess NDI reliability. Most of these programs had very similar objectives and methodologies. Over the years, the experience and technology also have refined various components of these attempts. This report reviews and evaluates these past NDI assessment programs. To understand the evolution at the program element level, the organization and presentation of the collected material are somewhat unconventional. First, the areas that contribute to the NDI reliability assessment (Examples listed below) were defined and listed as chapters for this report. Then major NDI programs over the past 30 years were studied and dissected so that their contents could be distributed among relevant element chapters. Chapter 2 provides a brief program overview on each of the major programs. Chapters 3 through 10 cover the various program elements. They are as follows:

- Facility sampling (Chapter 3)
- Inspector sampling (Chapter 4)
- Specimen Configuration (Chapter 5)
- Inspection Scheduling (Chapter 6)
- Inspections (Chapter 7)
- Data acquisition and handling (Chapter 8)

- Data analysis (Chapter 9)
- Human Factors (Chapter 10)

Within each chapter, various major attempts at POD assessment are presented, other reported literature that could be accessed is reviewed, and observations are recorded. This style of presentation compares and evaluates each of the program elements independently and forms the basis for a systems engineering approach to the protocol program development. At the end of each chapter, a statement on good practice is made based on Karta's understanding from the various published reports. The outcome from various programs is summarized in Chapter 11.

The subsections in chapters 2 through 11 are consistently titled and organized to help those readers who might prefer to read this report in terms of program by program. Such readers can read the report in the sequence 2.1, 3.1,..11.1; and 2.2, 3.2,…11.2; and so on.

Chapter 12 provides a brief on reliability programs not covered in chapters 2 through 11. Chapter 13 touches upon issues related to design of NDI reliability experiments. This chapter has value for readers who are interested in developing a reliability assessment experiment. Chapter 14 presents development efforts on computational and mathematical modeling of POD/ROC predictions. Chapter 15 reviews other related reports that could not be categorized in any of the previous chapters. Chapter 16 provides a summary of observations and list of most valuable resources identified during the review. Chapter 17 closes the report with our vision for a future NDI reliability assessment program.

Readers with very severe time constraints may read just chapters 2 and 16.

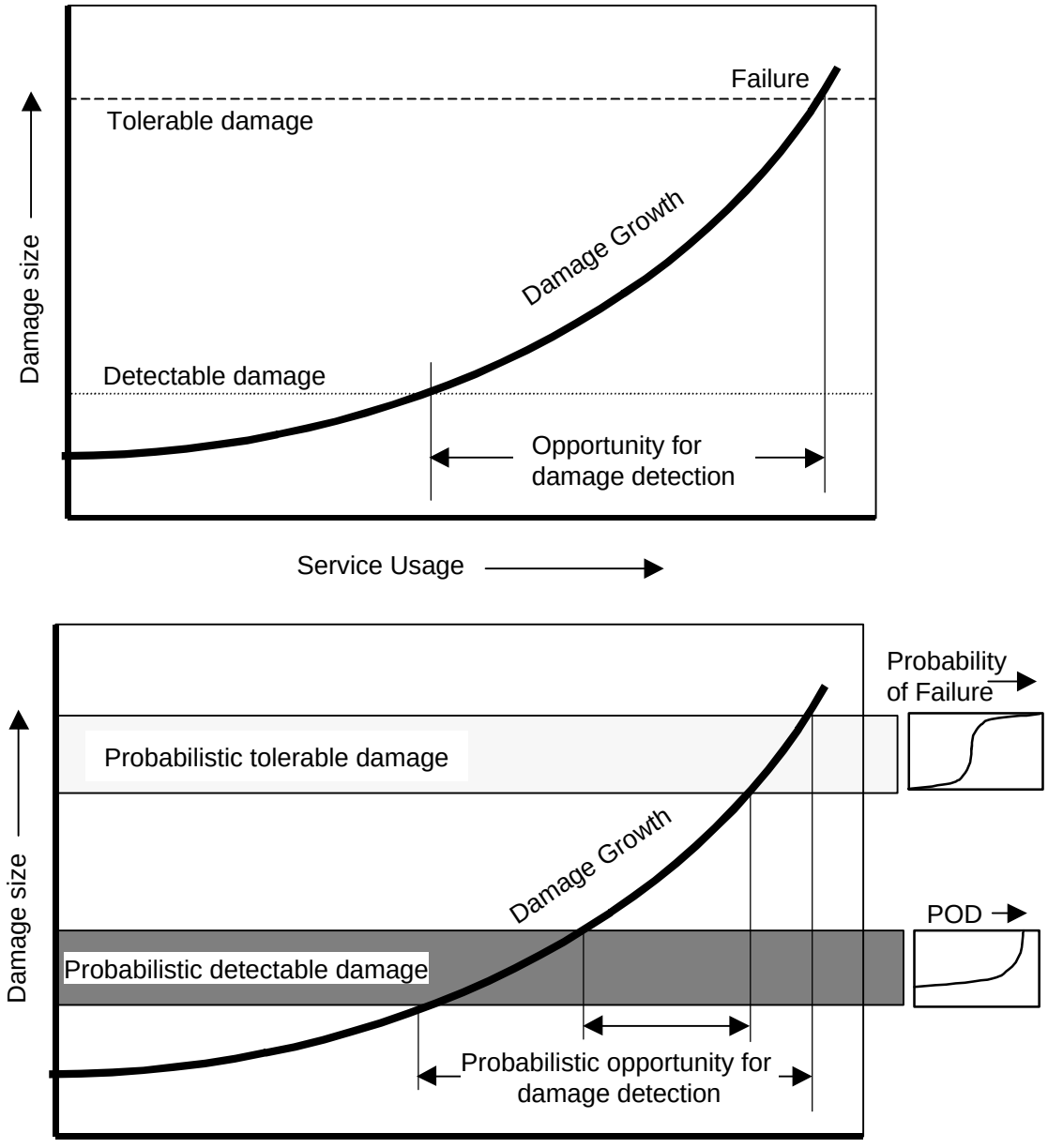

Figure 1.1 Damage Tolerance. *The above graphs qualitatively represents the concept of damage tolerance, where opportunity for damage detection is during the period when the damage grows from detectable level to failure level. Quantification of NDI capability and POD is important to obtain realistic estimate of damage growth life.*

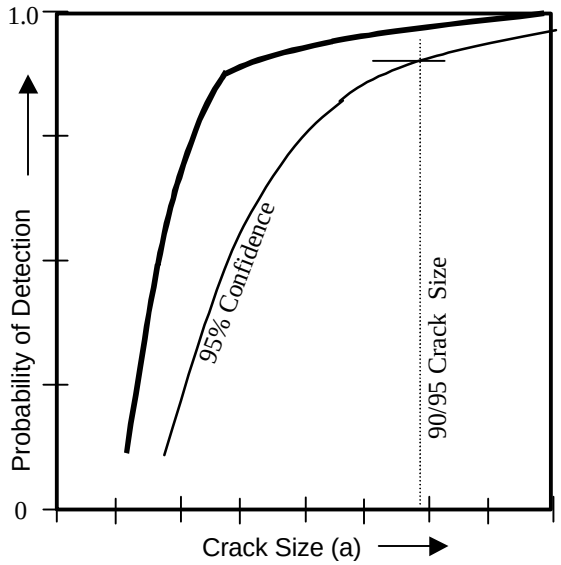

Figure 1.2 POD Curve. *The bold line represents the probability of detection variation with crack size. The thin line is the lower 95% confidence limit. A qualification metric can be a crack size for 90% POD with 95% confidence limit.*

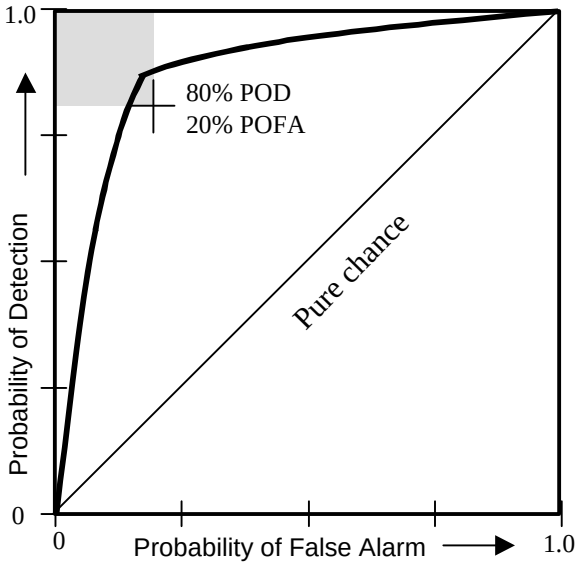

Figure 1.3 ROC Curve. *The bold line represents the relative operating characteristic curve. The thin line represents outcome of a pure chance. A qualification metric can be a minimum of 80% POD with a maximum of 20% POFA, or the area under the curve [Tanner 54].*

2. PROGRAM BRIEFS

The most common approach to POD measurement is to make specimens representing the actual part and introduce artificial flaws such as notches using Electronic Discharge Method (EDM) or fatigue cracks to simulate service-induced cracks. Although this approach is practical, in many situations it is not economical or possible to simulate the exact component or flaw geometry. Another approach could be to use the actual field inspection data and the crack growth curves to obtain POD. After detecting a crack in a part, the associated crack growth curve is used to estimate crack sizes missed in previous inspections. The approach is economical but requires collecting accurate and consistent data and recording over a long period of time. Alternatively one can use the actual parts that are known to contain flaws and perform NDI demonstrations using the same field inspections. This approach is both representative and practical; however flawed parts are not always available. A more recently evolving procedure could be a simulated inspection in a computer environment.

Significant work has been done by Martin Marietta, General Dynamics, Lockheed Georgia, Battelle, and Southwest Research Institute for the Air Force (AF); by Sandia National Labs for the Federal Aviation Administration (FAA); and Battelle for the Nuclear Regulatory Council (NRC). Universities, Electric Power Research Institute (EPRI), and various other laboratories have contributed to the science and engineering of NDI reliability assessment. Different programs supported by different organizations and industries had slightly varying objectives, but most of them have been philosophically similar.

This chapter provides a brief program overview of some of the major attempts at NDI characterization in the past three decades.

Evolution of NDI Reliability Assessment

"Aircraft Structural Integrity Program, Airplane Requirements" [MIL–STD–1530A], was in its formative stage in the latter half of the 1960s. The philosophy was changing from deterministic views with allowances for data scatter to probabilistic approaches for data acquisition, analyses, and modeling. Two essential elements referenced in MIL–STD–1530A are MIL-A-83444 (Airplane Damage Tolerance Requirements) and MIL-HDBK-5 (Metallic Materials and Elements for Aerospace Vehicle Structures). It followed that production inspection capabilities be statistically demonstrated, given initial flaw sizes and flaw growth modeling on fracture-critical parts. A big driver was loss of an F-111 on a training flight out of Nellis AFB. Failure cause was a fatigue crack originating from a very small production flaw in a D6-AC wing pivot fitting. Early work to statistically assess NDI was done by Packman, Pearson, Marchese, and Owens at Lockheed Georgia [Packman 68]. The AF then embarked on the program “Probability of Flaw Detection for Use in Fracture Control Plans” [Packman 76].

The concept of providing a statistical basis for quantifying NDI capabilities was formulated to meet requirements of the NASA Space Shuttle program. The NASA approach was to generate a large number of detection opportunities and assess the detection capabilities of various NDI procedures. Space Shuttle design and life cycle management was based on the results of this program. NASA produced flat panel specimens which contained 328 fatigue cracks, in a size range from 0.003"-0.750" in NASA Al2219-T6 alloy [Rummel 74]. The crack aspect ratio varied from 0.2 to 0.5. Panels were subjected to progressive liquid penetrant, UT, ET, and RT in the as-machined, after-etch, and post-proof conditions. Three separate operators during each inspection procedure performed NDI procedures. At the end of the assessments, all specimens were broken for crack characterization. Since the data was not sufficient to plot at the 95% confidence level (60 independent trials), it was analyzed at the 90% confidence level (29 trials) using a moving average method. The selected analysis scheme was to order the data from largest crack to the smallest crack, count down 29 observations, plot the point estimate of detection (finds/29) as a single point on POD curve, discard the largest crack in the first group, add the next smallest crack in sequence, repeat the point estimate, and plot the point estimates. A least square curve fitting analysis was used to plot a smooth curve. Since detection by some NDI methods is a function of both crack length and depth, NASA performed independent analysis of detection capability as a function of crack length and depth to envelope the crack detection capabilities of the four NDI procedures.

General Dynamics, San Diego conducted a parallel program and exchanged specimens with Martin Marietta to provide additional data sets [Anderson 73]. Both the programs used Marietta Marietta method of plotting and used the results as a basis for the design.

Martin Marietta conducted additional work to establish capabilities for detection in titanium flat plate; steel flat plate; aluminum weldments; and aluminum shapes. [Rummel 75, Rummel 76]

After establishing the capabilities of applied NDI methods, NASA validated the vendor qualifications. NASA used a subset, “29 out of 29” point estimate method to demonstrate that a vendor was performing at least to the level previously demonstrated by the full POD method [Rummel NY].

In the same time period, US Air Force Materials Laboratory conducted a program to provide a basis for design of the B-1 bomber. The Air Force produced a series of specimens containing flaws (fatigue cracks) of equal size at the assumed design size. Unfortunately, the production inspection processes were not capable of meeting the assumed design size and the test specimen cracks were not consistently detected. Failure to detect at an assumed flaw size provided no information on the actual capability of the production processes. The NASA POD method was thus adopted as an aid to improve NDI process improvements establish practical design limits.

Once the POD methodology was accepted, industry sought improvements reduce the amount of data required and to explore methods of extracting additional information. Two major programs were conducted with these goals: (1) NASA sponsored program

conducted by General Dynamics, Fort Worth explored various methods of data analysis using the NASA generated data as a basis for assessment and validation [Chang 76]; and (2) The Air Force Materials Laboratory sponsored program conducted by the University of Dayton Research Institute (UDRI) using Air Force ("Have Cracks") data [Berens 82].

The Berens–Hovey method proved useful in reducing the number of test specimens required and provided an added option for use of flaw size measurement in NDI capabilities analyses. The Berens–Hovey method has evolved as the most commonly used method and is the basis for the Military Standard for NDI capabilities assessment [MIL-HDBK-1823].

2.1 AF - Lockheed Georgia

1974-1978 Have Cracks Will Travel

Lockheed Georgia Company carried out a four-year study for the Air Force Logistics Command to determine the reliability of the AF NDI program [Lewis 78]. The objective of the program was to determine the capability of NDI to detect flaws under depot and field conditions. Over 300 Air Force technicians from 21 different Air Force Bases and depots performed over 800 ultrasonic testing (UT), eddy current testing (ET), penetrant testing (PT) and radiographic testing (RT) NDI tasks on actual aircraft samples containing fatigue damage. The program revealed that 90/95 reliability criteria could not be obtained for any flaw size with typical inspection techniques applied by the average technician. The program was nicknamed - "Have Cracks - Will Travel" or in short "Have Cracks".

This was the first major reliability assessment effort for NDI of airframe components under field conditions. All earlier efforts were directed towards capability assessment under laboratory and production conditions. The program generated enormous amounts of data that was also used much later (1980s) to validate improved data analysis procedures.

1979-84 NDI Technician Proficiency Program

The findings of the earlier program (Have Cracks) led to the development of a detailed test plan for NDI technician proficiency evaluation and documentation of activities necessary to conduct practical flaw-detection examinations with hardware [Lewis 80]. The test plan was limited to ET and UT methods, initially (1979-80). The approach was to generate fatigue cracks of varying lengths and depths at fastener sites in specimens, and mount the specimens on a rack to simulate the piece of a structure. The NDI technicians were provided with detailed procedures and asked to search for flaws (1980-82) [Sproat 82]. A database of their responses formed the basis for establishing performance norms. This database was further used in (1983-84) to evaluate improvements in POD in comparison with the "Have Cracks" program. The attention

paid to NDI technician proficiency following the "Have Cracks" program produced improvements in USAF NDI reliability. However, the detection probabilities for some NDI methods were still below the levels established as standards for structural analysis by MIL-A-83444. The program stated "continued effort to improve POD and reliability for the NDI methods in use through better technician performance, better equipment, and better procedures seems the most effective way to achieve the needed POD and reliability levels" [Sproat 84].

1987-88 Engineering Services in Support of NDI

This program addressed NDI capability and reliability measurements on a number of flaw types and hardware configurations beyond the scope of the 1974-78 program cited above. Composites were treated along with several other metallic structural configurations [Sproat 88]. The work was co-performed by UDRI.

2.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Martin Marietta conducted a four-year study on NDI reliability of engine components during overhaul at Kelly and Tinker Air Force Base (AFB) [Rummel 84]. The assessment program entailed the collection, flawing, and characterization of gas turbine engine components in accordance with the plan, conducted on-site NDI assessment using production NDI procedures, documented the response and analyzed the data to establish the baseline capabilities and reliability of each NDI procedure. The project supplemented engine components with test specimens to enable assessment of specific areas of the components that could not be addressed directly. Simultaneously, project team performed an off-line assessment of fluorescent penetrant processing materials and correlated it to the POD analysis. They identified major variations in NDI engineering and in process application and control and lesser variations in human factors associated with process applications.

This was the first major reliability assessment effort for NDI of engine components in production-line environment. Once again, the overall NDI reliability in engine overhaul was below that which had been assumed or generally desired.

1984-1989 Assessment of Automated and Semi-Automated NDI Processes/Systems

Martin Marietta conducted additional assessments on gas turbine engine overhaul lines to assess various aspects and capabilities of both automated and semi-automated NDI procedures. These included: assessments of penetrant processing lines [Rummel 82a], assessment of automated and semi-automated eddy current inspections [Rummel 86], automated penetrant processing of engine blades (IBIS) [Rummel 86a], and penetrant materials process quality control/measurement methods [Rummel 82b]. Martin Marietta

also conducted assessment and revalidation of NASA contractor capabilities in parallel [Rummel 83, Christner 88].

Automation can resolve some of the human factors and improve NDI reliability, but the NDI engineering task is equally critical. Variations in inspection materials, equipment, processes, and personnel necessitate validation of baseline performance and periodic assessment of performance for individual inspection processing lines.

2.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

The objective of this project was to develop an NDI process assurance methodology for AF NDI labs [Hyatt 88a]. Battelle conducted a series of visits to selected NDI field labs. Then they developed NDI process assurance methodologies that seemed consistent with the personnel, facilities, capabilities, and activities of the field labs and simultaneously meet the process assurance needs for the generation of quantitative performance-based information. This work culminated in selection of the preferred approach that was subsequently tried in a single lab field trial. Finally they conducted a field trial involving a number of operational NDI field labs.

2.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

This one-year project resulted from earlier findings in regards to NDI technician proficiency levels. The goal of this effort was to recommend a program for increasing AF NDI technician proficiency. The approach taken was to (1) identify relevant areas of concern that could negatively impact the proficiency of AF NDI technicians; (2) seek possible solutions for identified concerns; and (3) blend the most feasible, promising, and cost-effective potential solutions into recommendations for improving proficiency [Schroeder 88].

SwRI used four sources of information: literature dealing with AF NDI technician proficiency; literature dealing with industrial NDI technician proficiency; comments generated by AF personnel during interviews and other coordination activities; and comments generated by SwRI investigators from observations, interviews, site visits, and other coordination activities during the course of the project.

1986-88 NDI Personnel Proficiency Evaluation Using UT

The project involved testing the proficiency of NDI technicians at the San Antonio Air Logistics Center (SA-ALC) in the inspection of F-100 stage 2 compressor blades, using

the UT method. The results could be used to establish general training program requirements and also the NDT capability of an individual inspector [Davis 88].

1990-94 Reliability Assessment Kit

Subsequent to the recommendations on technician proficiency improvement, SwRI contracted with the USAF to develop technology and representative test samples necessary to assess the capability of USAF technicians to nondestructively evaluate USAF airframes [Goodlin 94]. The newly defined objectives were to (1) develop a statistically significant methodology for testing USAF technicians based on a test plan previously developed by Lockheed Aeronautical Systems Company, (2) provide a standardized NDI kit and use it to test the technicians at Air Force and Naval NDI Labs, and (3) reduce and analyze the data gathered from testing the technicians in five NDI disciplines. The SwRI developed test kits comprised of 17 standardized NDI procedures and one software procedure.

2.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

The FAA funded a team comprising of Sandia National Labs and Science Application International Corporation (SAIC) to experimentally assess the reliability of high-frequency eddy current inspections of lap splice joints in aircraft maintenance and inspection facilities [Spencer 94, Ashbaugh 95]. Trained monitors traveled with the experiment and recorded not only the inspection results, but also observations on the maintenance environment and procedures. The team used 13 well-defined protocols during the experiment. They simulated the lap splice inspection task using test specimens with known and well-characterized crack lengths. A set of five inspections was performed at nine facilities, in an actual environment. Substantial variation occurred from inspector to inspector and facility to facility. Areas identified for improving field results were: (1) getting inspectors to pay closer attention to procedures as well as optimization of procedures and (2) ensuring that inspectors were better trained regarding their specific equipment. The need to investigate lap joints came out of the Aloha accident in 1988, which led FAA to sponsor a major program on aging airplanes.

2.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

SAIC developed and validated a UT Procedure to inspect for fatigue cracks in the 1st and 2nd layers in the C-141 lower wing spanwise splice joint [SAIC 98]. The inspection procedure incorporated automatic scanning of the inspection surface and automatic

saving of the data in electronic form. The impact of the human inspector on the reliability of the process was only in the setup of the equipment and interpretation of the data. SAIC designed the methodology to achieve three key objectives: (1) determine the inspection reliability of UT, (2) identify the areas where improvements in reliability could be made, and (3) assure that high reliability could be maintained with the existing line of inspectors. The SAIC team followed the protocol developed by Sandia National Labs for FAA [Spencer 93]. For the first time, the Design of Experiments (DOE) approach was applied to study the influence of various variables on NDI reliability assessment.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

Following successful completion of C-141 lower inner wing spanwise splice second layer inspection process, SAIC identified other areas in need of inspection procedures. Next, SAIC incorporated the capability of 1st layer inspections running concurrently with the 2nd layer inspection process [SAIC 99]. This combination was accomplished with equipment modifications. The small-scale validation made use of as much of the equipment and personnel from the original POD study as possible.

1998-99 C-130 Center Wing Stringer - Hat Section

This study ran on lines parallel to the second layer POD study for C-141. Its purpose was to investigate the impact of change in application.

2.7 NRC-Battelle

Various organizations, such as EPRI, ASME, NRC have conducted various programs to support the nuclear industry. While there is commonality of some inspection equipment, methods and inspection materials, both the test problems and the requirements differ significantly. With limited commonality in NDI methods, few results and little data from the nuclear programs are applicable to the aerospace industry.

1985-86 Mini Round Robin Assessment of UT Performance

The Nuclear Regulatory Commission (NRC) sponsored a Mini Round Robin assessment at Pacific Northwest Laboratory (PNL) to evaluate the ability of NDI technicians to detect intergranular stress corrosion cracking (IGSCC) using UT [Wheeler 86]. Battelle concurrently conducted a limited human factors study to acquire preliminary data on performance-shaping factors that effected UT reliability and tested the efficacy of ROC analysis for representing UT accuracy. Technicians performed UT on welded piping sections, filled out a questionnaire, and were interviewed on inservice inspection experience. In addition, Battelle evaluated UT equipment for conformance to human factors design principles.

1985-86 Human Reliability Impact

Human factors research in areas related to NDI (medical and industrial) and human factors research in NDT were evaluated to determine what classes of variables would be most likely to affect NDT performance [Triggs 86]. Battelle studied task, procedural, training, individual difference, and environmental variables. A secondary purpose of this project was to develop a satisfactory measure of performance that could be applied to NDT.

2.8 MIL-STD-1823 (Draft)[2]

MIL-STD-1823 (draft) provides the requirements and methods for test and evaluation procedures for assessing NDI system capability. This document can be used to demonstrate that an NDI system can meet specified requirements and also to identify and measure major sources of variations. It addresses ET, PT, UT and MT only; however, the document may be used for other NDI procedures if they are similar in output. It classifies NDI systems into two categories: those that produce only qualitative information (i.e., hit/miss data) and those that provide quantitative information as well (e.g., signal amplitude). The main document addresses general requirements and responsibilities for planning, conducting, analyzing, and reporting NDI reliability evaluations. Specific requirements for each of the NDI techniques are put in appendices.

The document presents very valuable information and guidelines that are easy to understand and follow. A number of recommendations are still valid and can probably be used for forthcoming programs. The technological developments in the area of numerical simulation and data processing can add value to the basic procedures presented.

2.9 Other Published Work

Scientific literature reports a number of other minor NDI reliability attempts. Some of these are separately reviewed in Chapter 12.

2.10 Observations

- The subject of NDI capability and reliability assessment started drawing serious attention from an operational viewpoint in late 60’s and early 70's.

[2] This document remained a draft for 10 years (1989-1999). During this period, it was used and referred to many times. Throughout our review, we cover this Draft rather than a recently accepted MIL-HDBK-1823, which is very similar to the original draft.

- Initial assessment programs with the Air Force on aircraft and engine components revealed deficiencies in detection reliability.
- Subsequent efforts, including training, skill development, and re-evaluations have improved technician proficiency, but continuous improvement is needed.
- Programs have not reported much evidence of usage of experience from previous programs, except that "Have Cracks" and MIL-STD-1823 (draft) have been cited quite often.
- No major program has been directed to assessment of the results of incorporating recommendations made or documentation of the results of implementation.
- AF-Battelle, AF-SwRI, and development of MIL-STD programs were aimed at providing AF with guidelines and tools for continuous NDI reliability assessment on a sustained basis. Battelle and SwRI programs met their objectives; however, no evidence of their continuous use could be seen. MIL-STD-1823 (draft) is a good document with very useful set of guidelines on design and execution of reliability experiments, and has only recently been accepted as MIL-HDBK-1823. Many of their recommendations are still valid and correct. They should be used in future programs. The details are discussed in subsequent chapters.
- Most programs aimed at reliability assessment. No report attempted an assessment of what best can be achieved (peak performance prediction or capability assessment).
- Most programs were dominated by field experiments. Computer simulations and modeling can supplement the experimental data.
- A lack of good human factors investigation and mitigation program still exists within the aviation industry. The nuclear sector demonstrates that this subject lacks useful guidelines.
- Human factors are often cited as the primary avenue for improvements in capabilities. Unfortunately, many of the human factors programs have been focused on assessment of factors within existing NDI process applications rather than on improvements in the tools and other process improvements that would aid the operator in effectively carrying out an NDI procedure. There is little evidence that action has been take to implement recommendations made for those factors that have been identified as significant contributors by various assessment programs.

Good Approach: AF-SAIC

3. FACILITY SAMPLING

Synonym usage for Facility: Base, Depot, Lab.

To obtain an assessment of overall NDI capability, the Air Force requires an average reliability outcome of all inspections that they conduct. Ideally all facilities, depots, labs, and bases should be evaluated; but economics of the program limit the experiment to a few sampled locations. Different programs visited different facilities for various reasons. This chapter documents the bases touched and possible reasons for their selection to participate in the program.

3.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

This program visited the following facilities:

AFC	Hill, Kelly, McClellan, Tinker, Warner Robins
ATC	Randolph, Reese, Webb, Williams
MAC	Charleston, Dover, McChord, Travis
SAC	Carswell, Ellsworth, Offutt, Pease
TAC	Bergstrom, George, MacDill, Shaw

These sites were selected to cover all different types of maintenance facilities, within geographical constraints of travel, in order to provide the best possible average across the USAF.

1979-84 NDI Technician Proficiency Program

Lockheed conducted the initial study at Dover, which helped determine the minimum and optimum number of inspection sites and flaw detection opportunities necessary to conduct a valid statistical measure of practical NDI performance. They later conducted evaluations at other AF bases.[3]

1987-88 Engineering Services in Support of NDI

For this project, Lockheed performed facility sampling at two levels. The AF was to decide which of the commands to include in the evaluation program after reviewing the project's final report. The second level determined that three to five bases in each command should be included. The optimum sampling allocation with the main goal of

[3] The report on further results of the program could not be obtained

estimating the overall AF NDI detection capability called for dividing the total number of bases visited proportionally to inspection activity at each command, with a minimum of three bases in each command. The Alaskan Air Command, Pacific Air Command, and the USAF in Europe were not considered because their geographic location were outside the continental US, which could make the operations expensive. The chance of a base being included in the evaluation program should be proportional to the volume of inspections conducted at that base. Lockheed used a randomized computer selection process with different probabilities associated with different bases.

3.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Martin Marietta identified engine overhaul facilities at SA-ALC, Kelly AFB, and OC-ALC, Tinker AFB for NDI capability and reliability assessment in terms of "fracture control" or "retirement for cause philosophy." These were the only engine overhaul facilities.

3.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

During the spring of 1987, Battelle visited six NDI field labs. They were Williams (ATC), Luke (TAC), Nellis (TAC), Beale (SAC), Travis (MAC), and Mather (ATC). Battelle selected these labs because they represented a broad spectrum of aircraft and commands, within close geographical proximity to each other. In the second round, the four labs visited were Alconbury, Leakenheath, Bentwaters, and Upper Heyford. These four were selected to coincide with the sites for the fall 1987 field survey. At the end, Battelle visited RAF NDI lab at Brize-Norton because it was in the general area at the completion of field trials.

3.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

SwRI visited Laughlin and Chanute AFB to discuss issues with the NDI personnel. The reasons for this selection are not recorded in the report.

1986-88 NDI Personnel Proficiency Evaluation Using UT

SwRI conducted UT proficiency testing at SA-ALC only. The reason for this selection appears to be geographic proximity of SwRI to SA-ALC.

1990-94 Reliability Assessment Kit

SwRI conducted the validation of the reliability assessment kit at Kelly AFB. The reason for this selection appears to be geographic proximity, and the AF NDI office was at Kelly at that time.

3.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

Sandia and SAIC took the experiment to nine facilities. They chose the facilities to obtain a cross section of those where inspections of transport aircraft were performed. Major attributes considered were size of the inspection force and in-house versus third-party inspections. The facilities visited were American Airlines in Tulsa, OK; Dalfort Aviation in Dallas, TX; Aloha Airlines in Honolulu, HI; Tramco in Everett, WA; Alaska Airlines in Seattle, WA; United Airlines in San Francisco, CA; Delta Airlines in Atlanta, GA; US Air in Winston-Salem, NC; and Miami NDT in Opa-Locka, FL. The cross section covered large, small and third-party inspection facilities, and represented the conditions under which typical inspections were performed. Facilities were not under test.

3.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

SAIC's facility in New London, Connecticut conducted the laboratory validation of the inspection procedure. SAIC performed the field validation and actual tests at WR-ALC.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

This study was conducted like the previous study.

3.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

Pacific Northwest Laboratory performed the round robin assessment.

3.8 MIL-STD-1823 (Draft)

This document does not mention facility sampling.

3.9 Other Published Work

Nothing relevant was found.

3.10 Observations

- In general, facility sampling appears to have been a matter of logical choice, rather than a scientific activity.
- To verify or calibrate an assessment procedure, the most conveniently located facility was generally selected; and that is probably a good practice.
- Lockheed's approach “the chance of a base being included in the evaluation program should be proportional to the volume of inspections conducted at that base”, is a good approach.
- The total sample size for an AF-wide evaluation must cover different types of facilities (labs and depots, large and small).

Good Practice: AF-Lockheed

4. INSPECTOR SAMPLING

Synonyms usage for Inspector: Technician, Practitioner, Operator

To obtain a good, dependable performance for the AF in the field or at a facility, all inspectors should be asked to go through NDI evaluation. Once gain, economics, time, and availability constraints demand that inspectors be sampled for the objective. The optimum method for sampling depends on the goals of the evaluation program. Some sampling plans are geared towards providing precise estimates of the overall population mean, while others are better for calculating estimates of means of sub-populations. An important assumption for the validity of statistical analysis is that the sampling be performed in an unbiased and random manner.

4.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

Lockheed program had an average of 15 participants at each ALC and 6 participants at each field-level base within the commands visited. The program plan made no attempt to prevent marginally performing technicians, nor equipment with minimal capabilities, from participating in the program. AF on-site management assigned both technicians and equipment on an availability basis irrespective of expected performance. The participant selection from those available was essentially random. Some bases had a limited number of personnel; and, at those bases all technicians participated in the program.

1979-84 NDI Technician Proficiency Program

Theoretically, the program-developed test plan could be employed for proficiency evaluation of any technician. The total number of technicians tested through this program were 121 for UT, 133 for ET, 84 for RT, 37 for PT, and 29 for MT.

1987-88 Engineering Services in Support of NDI

The Lockheed program recommended as many inspectors as economically feasible with at least five for each inspection method at each base. The selection of inspectors did not need to be a mutually exclusive group of inspectors. A simple lottery selected a sample of inspectors, with each qualified inspector having an equal chance for selection.

4.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

The goal of the program was to test 15% of the certified inspectors at a facility and 10% repeat inspections by a single inspector. At each facility, the program selected inspectors at random from the certification list.

4.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

For a nominal schedule to cover five NDI methods, Battelle selected a probable sample size of five NDI practitioners at random from the labs assigned-duty staff. Since the lab size varied between five and twenty, this sample size represented a significant fraction of the total population. This meant that the effect of sampling was consistent for a statistical analysis of the data.

For a hyper-geometric distribution, the expression for the probability of exactly x proficient practitioners being in a sample of n practitioners, drawn at random from a population of N practitioners that contains X proficient ones, is given as follows:

$$f(x;n)=\frac{\binom{X}{x}\binom{N-n}{X-x}}{\binom{N}{n}} \quad ; \qquad \text{where} \quad \binom{\mathrm{i}}{\mathrm{j}}=\frac{i!}{j!(i-j)!} \qquad \text{and} \quad ! \quad \text{indicates factorial.}$$

Similarly, the expression for the probability of x or more proficient practitioners being in such a sample is given as follows:

$$F(x,x+1,\ldots,n;n)=\sum_{i=x}^{n} f(i;n)$$

These formulae can help decide on the number of practitioners that should be sampled for a reliability experiment; however, initial judgment of the number of proficient practitioners N is always subjective.

4.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

Not applicable.

1986-88 NDI Personnel Proficiency Evaluation Using UT

The subjects were the NDT technicians of the Structural Assessment Testing Facility at SA-ALC, Kelly AFB, Texas. Phase I involved 24 technicians, out of which 11 were re-tested in Phase II.

1990-94 Reliability Assessment Kit

All SA-ALC inspectors took the validation test since the number of inspectors was small. SwRI involved at least 15 SA-ALC technicians in validating kits for each of the five NDI methods.

4.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

At each of nine selected facilities, four inspectors (or inspection teams) completed the inspection tasks. At each facility, one of the four inspectors performed an inspection a second time. The net result was 45 inspections. The selection of inspectors was left to the managers of the facilities visited. In some cases all the inspectors participated.

4.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

Fourteen inspectors participated in C-141 lower wing spanwise splice joint 2nd layer cracking. The inspectors included six experienced equipment operators with only procedure-specific training, five AF NDI technicians with no UT scanning imaging equipment experience but with extensive equipment operation training plus procedural training, and three AF NDI technicians with limited UT scanning/imaging equipment experience who received equipment operation refresher and procedural training.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

SAIC selected five inspectors from those who participated in the previous program or who were currently using or trained in the second layer inspection process. Two of them were initially identified as having no prior experience with Ultra Image IV or similar equipment. Three others were familiar with the operations but had no production experience. All of these inspectors received training on the new procedures.

4.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

A total of 12 UT technicians participated in the program, two each from six organizations actively involved in UT inservice inspections and certified at level II or Level III. Each participant had successfully passed the IGSCC detection performance demonstration at the EPRI-NDE Center in Charlotte, NC. Actual selection was made by each company, which was assumed to be made on availability and the individual's desire. The technician age range was 27-50 years with an average age of 35, their experience range was 1-12 years with an average of 7.4 years, and their formal NDT training range was 1-74 weeks with an average of 16 weeks.

4.8 MIL-STD-1823 (Draft)

MIL-STD-1823 (Draft) recommends that test plans should include participation of several inspectors selected at random from among the eligible population. Eligibility may be defined in terms of some particular certification, training, or physical ability.

4.9 Other Published Work

Not applicable.

4.10 Observations

- Most of the programs selected a certain minimum number of inspectors at random from those available or eligible.
- Bias is possible if inspectors are to be selected from those available. A supervisor worried about his facility being evaluated may offer the best of his inspectors, and a supervisor concerned about the production line may send those inspectors who are not as good. In earlier days supervisors rarely knew the identity of the highly capable inspectors; however, over the years, various inspection proficiency tests have revealed the individual capabilities.
- A certain minimum number (say 5) or a percentage (say 15%) is a good choice. This, of course, may require all inspectors to participate from a smaller facility for a meaningful sampling.
- Hypergeometric distribution expressions can help decide on the number of practitioners that should be sampled for a reliability experiment at each facility.

Best Practice: AF-Battelle

5. SPECIMEN CONFIGURATION

Synonyms for Specimen: Sample, Standard

Synonyms for Inspection site: Inspection Opportunity, Operator Judgment

Specimens are the heart of the NDI reliability assessment program. It is very important to have a statistically significant number of well-characterized flaws in a set of realistic structural components presented to the inspector in a natural manner that provides no clues as to presence or absence of a flaw.

5.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

The Lockheed program used six different types of actual aircraft structure with fatigue cracks as specimens: (1) Type A: One C-130 center wing box, intact, five foot single piece; (2) Type B: Twelve C-130 center wing box lower-surface segments; (3) Type C: Simulated titanium wing risers mounted inside Type A; (4) Type D: Two sections of KC-135 center wing lower plank; (5) Type E: Two F-104 wing fitting segments; and (6) Type F: One C-5 wing spar cap/web test assembly (box beam). Part of the specimen set was from another exploratory work to simulate in-service NDI using C-130 center wing boxes removed from service, fatigue cycled to generate and document multi-site damage, and statically loaded to determine residual strength at failure.

There were a total of 238 cracks in the set of specimens with sizes varying from 0.01 to 1.05 inches in the total program. These specimens and the inspection program gave approximately 46000 measurements.

During the program, 76 technician proficiency-screening samples were added in Jan. 1977 to determine if a relatively simple small inspection sample could be used to evaluate or predict technician NDI ability on larger more complex structure. These structures were 2x16x0.20 inches bare Al alloy sheets with 10 holes each with randomly distributed fatigue cracks at 123 holes.

1979-84 NDI Technician Proficiency Program

The initial Lockheed study at Dover AFB recommended 138 fastener sites for eddy current NDI and 108 test sites for ultrasonic NDI for fatigue crack detection at fastener holes so that 90% of the participants could complete the test in a regular eight hour day. The participants performed the ultrasonic test on a racked assembly of aluminum elements, which simulated a wing spar. Fatigue cracks radiated from fastener holes in the web and the cap. Participants conducted eddy current bolt hole NDI on a similar assembly with fasteners removed.

The participants also performed X-ray radiography on a simulated aluminum skin stringer panel with fatigue cracks radiating from fastener holes in sandwiched shims. Fluorescent penetrant test hardware consisted of racked titanium elements simulating riveted panels where fatigue cracks extended from fastener holes. The test hardware for magnetic particle NDI consisted of steel clevis/link rod assemblies with fatigue cracks at clevis roots and holes. The link rods contained flaws, which were an exception to the fatigue crack design. These rods contained EDM notches at thread roots and mid section.

The modular hardware system consisted of a rack, which could accommodate three types of interchangeable specimens. All contained the same spectrum of flaws, but allowed for different tests among individuals.

1987-88 Engineering Services in Support of NDI

Under this program Lockheed developed six standard sets.

Fastened joint standard consisted of four different racks, each with a set of four skin and J-stringer splice with two rows of fasteners. Material was Al 7075-T6 with no surface finish. Fatigue cracks of varying lengths emanated from fastener holes.

Lug standard consisted of a single lug joint sandwich welded between two angle plates. Twenty seven lugs of 4340 steel alloy with a polyurethane primer coat were used for magnetic particle NDI. Fifteen lugs of 6061-T6 Al alloy with a polyurethane primer coat were used for eddy current NDI. Cracks of varying lengths extended radially from some holes and longitudinally in the lug attach angle root.

Flat plate standard consisted of three 7075-T6 Al rack assemblies, with 16 Ti-6Al-4V plates with no surface finish per rack and four sites per plate. Fatigue cracks of varying lengths emanated from plate edges.

Honeycomb standard consisted of honeycomb core with graphite/epoxy and kevlar face plates without surface finish. Four specimens, each 5x18 inches were subdivided into six areas separated by milled grooves. Damage included (1) impact damage in face sheets 1/4 to 3/4 inch diameter circular, (2) delamination flaws in face sheets from 1/8x1/4 inch rectangular to 3/4 inch diameter circular, and (3) Detached core flaws at the core-face bondline from 1/8x1/4 rectangular to 3/4 inch diameter circular.

Fastened transition joint standard consisted of graphite/epoxy laminates fastened to 7075-T6 Al base plate without any surface finish. Four transition joint standards 5x18 inch were subdivided into six areas separated by milled grooves. Damages were introduced around fastener sites to simulate hole wear in composites. The damage size varied from 1/4 to 1 inch.

Bonded transition joint standard consisted of graphite/epoxy laminates bonded to 7075-T6 Al base plate without any surface finish. Four transition joint standards 5x18 inches were subdivided into six areas separated by milled grooves. Delamination flaws were introduced ranging from 1/8x1/4 inch rectangular to 3/4 inch diameter circular within various laminae and at the bond between the plates.

For all standards, all configurations contained the same number of inspection sites, number of flaws, and same range of flaw sizes. Specimen standards were identified with metal tags bonded at the upper right corner. When fully assembled, the identification tags were concealed.

5.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Martin Marietta program used applicable engine components as specimens. They supplemented the set with artificially flawed specimens. Test specimen materials were generic depending upon the component under considerations. The selection was Ti for fan sections (6Al-4V for specimens), Ni base for turbine sections (Inconel 718 for specimens), and low alloy steel for compressors and turbine shafts (4340 for specimens). The program recommended 10 sample test sets with a recommended minimum of 100 flaws per test set. The actual distribution of flaws was obtained using the tear-down inspection. The program also recommended at least 50% unflawed specimens. The general distribution of flaw lengths was 0.010 to 0.500 inch.

The program of specimen selection began with procurement of a large quantity of scrap TF-30 blades, vanes, and disks. A thorough examination of these scrap components revealed countable flaws. It was concluded that either the flaws were not detectable or the number of flawed components was small. SA-ALC and OC-ALC then verified a low rejection rate for flaws. ALCs were then requested to save all flawed components rejected for cracks by all NDI inspection methods. This yielded components with known flaws for some test sets. The hardware included Ni and Fe based turbine disks, Fe turbine blades, Ti fan blades, Ti fan disks. Disks were selected from those that were available in sufficient quantity to afford selection of a test set containing cracks in the size and population range desired.

For some configurations the number of parts were not adequate to support a full test set. Extensive effort was made to grow flaws in such configurations. It turned out to be a project in itself. Test sets were produced in Ni base turbine blades, Ti fan disk scallops, IN100 turbine disk scallops, and Fe base turbine blades. An additional test set was produced for ultrasonic surface wave inspection of Ti fan blades.

A total of 12 sets were produced at the program outset. During the NDI assessment, two of the sets were destroyed by repeated cleaning procedure. One of these was replaced by a 13^{th} set.

5.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

Battelle determined that they needed samples with a range of defect sizes for each of the five primary NDI methods. Based on the prior field trial experience at Pease AFB, Battelle designed these samples such that an inspector with minimal proficiency could find at least some of the defects whereas a considerable greater level of proficiency would be required to find all the likely detectable cracks. In addition the need for transportability, moderate cost imposed additional constraints on the samples selected. Samples finally selected were a combination of samples from the NDI technician proficiency kits, the Rummel engine inspection study, and those fabricated especially for this field trial study. An industrial-grade under-the-seat-brief case was configured to hold the specimens. The set for five methods included:

PT Samples: 10 coupons 1x5x3/32 inch with 2 3/16 inch diameter holes from test set #13 of Martin Marietta. Out of these 20 opportunities[4], 7 were fatigue cracked with 0.02-0.118 inch crack lengths.

MT Samples: 10 samples from J59 fan blades from test set #7 of Martin Marietta. Out of the 14 opportunities, there were 7 cracks with lengths verified to be between 0.033 to 0.383 inch.

RT Samples: from technician proficiency kit fabricated and validated by Lockheed. Out of the 12 opportunities at fasteners, 4 were flaw sites with crack lengths between 0.25 to 0.30 inch.

ET Samples: also from the technician proficiency kit fabricated and validated by Lockheed. This sample represented an aircraft skin section and contained cracks originating at the fastener holes. Out of these 10 opportunities, 3 were flaw sites with crack lengths from 0.125 to 0.25 inch.

UT Samples: specially fabricated from Aluminum blocks of 2x6 inch with EDM notches between 1/8 to 3/4 inch.

5.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

Not applicable.

[4] This is Battelle terminology for inspection site.

1986-88 NDI Personnel Proficiency Evaluation Using UT

The test specimens consisted of both flawed (artificially induced fatigue cracks) and good F-100 stage 2 compressor blades. The specimen set in Phase I had 38 unflawed blades, 56 flawed specimens with crack size ranging from 0.050 to 0.173 inch, and 56 flawed specimens with crack size below 0.050 inch, which was the limitation placed by the minimum flaw size in the test blades imposed by the inspection technique of the Technical Order (T.O.). Thus Phase I had two sections - full set of 150 specimens and a subset of 94 specimens. The specimen set in Phase II had 38 unflawed blades and different set of 12 flawed specimens with crack sizes ranging from 0.050 to 0.084 inch.

1990-94 Reliability Assessment Kit

The radiography testing kit hardware consisted of four mounting rack frames with four NDI standards assembled in each. Each standard had two angles connected to form a J stringer. One leg of the angle was attached to the base sheet with two shims underneath. Standards were designed with two rows of six fasteners equally spaced through the shims. These standards were installed in a frame. The standards and mounting racks were fabricated from aluminum sheets and angles. Fatigue cracks of varying lengths were introduced at some fastener holes in the shims extending towards the edge of the shims.

The specimens for bolt hole ET were similar to the RT. The ET surface examination kit hardware consisted of fifteen aluminum-lug ET standards assembled in three racks, with freedom to alter the sequence. The fatigue cracks were grown from the lug holes towards the edge of the specimen, at eight possible locations on each lug hole.

For PT, the test parts were flat plates fabricated from Ti-6AL-4V with a ground face. Fatigue cracks of specific lengths were grown from the edges, at four possible locations. For inspection, sixteen of these plates were mounted in an aluminum frame, and a total of three frames to be inspected.

For MT, the test parts were lugs fabricated from 4340 steel and a 4340 angle welded to the bottom. Fatigue cracks of specific lengths were grown from some bolt holes and root radii of the angles, at eight possible locations. After fabrication and flaw growth, each of the lugs was coated with two coats of polyurethane primer.

Honeycomb specimens for UT were fabricated from one inch aluminum honeycomb core with 12 plies of graphite-epoxy face sheet bonded to one surface and a 12-ply kevlar face sheet bonded to the other. Delaminations were introduced in the panel. Bonded joint transition specimens for UT were fabricated from graphite epoxy laminates, which were adhesively bonded to an aluminum 7075-T6 alloy base plate. Each test standard was 5x18 inch section subdivided into six areas or segments separated by grooves. Three fasteners were located in each segment. Delamination and disbonds of 1/8 by 1/4 inch to 1-inch diameter were introduced as flaws. A similar setup with a fastener joint instead of a bonded joint formed another procedure for UT evaluation.

5.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

Sandia and SAIC designed and fabricated the experimental hardware to simulate the fuselage of a typical narrow body transport aircraft. They engineered known flaws into two types of test specimens described below.

Type 1: 20x20 inch flat skin panels with two plates fastened together using three rows of rivets. These were assembled on a frame butted against each other. These pieces provided freedom to alter the presentation to each inspector, thus avoiding transfer of crack pattern knowledge from one inspector to another. Fatigue cracks were introduced at rivet holes. Forty three specimens were fabricated including seven as a backup in case of field damage to the main 36. The crack lengths of 172 cracks on 122 rivet sites were measured using SAIC Ultra Image and were found to be up to 300 mil. Of these 122 fielded rivet sites, 75 contained horizontal cracks, 21 were at 11^{o} angle, and 26 were at 22^{o} angle with respect to the horizontal. For the first four facilities the specimens were bare; and for the subsequent five, they were painted.

Type 2: Two 8x4 foot curved panels (75 inch radius) with a longitudinal lap splice simulated complete aircraft structure. Foster Miller fabricated these panels. Fatigue cracks were generated in the panel using a custom designed load machine that simulates bi-axial loading. Video micrometer system was used to characterize the skin panel. Cracks beneath the rivet head could not be seen, so the 83 cracks observed were in a 60-200 mil range. A complete characterization could have been obtained through sacrificing these specimens; however, it was decided that specimens could be of more value intact for future validation work. One of these specimens was painted and one was kept bare.

The team presented the specimens to the inspectors in a manner that could identify the impact of crack density on reliability. Each inspection row started with two unflawed panels. The remaining sixteen specimens were split evenly into an area containing approximately 10% flawed sites and an area containing 40% flawed sites. This was done to reduce the expectations of the inspectors.

5.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

The specimens were actual spanwise joint segments cut from C-141 A/C 66-0186 as part of Lockheed teardown inspection. The segments covered the entire lower inner wing spanwise joint thickness range. The segments were approximately 40 inches in length, and contained approximately 35 spanwise fastener sites each. Thirteen out of sixteen splice joint segments were subjected to cyclic fatigue testing with initial starter notches in second layer. At 59 sites the cracks propagated to cover a crack length of 0.030 to 0.250

inch. At three locations, cracks had propagated to the edge. The uncracked to cracked fastener site ratio was 4:1. The size, location, and orientation of the cracks was chosen using a random number function in Microsoft Excel. Each crack was completely characterized using high magnification optics, contact ultrasonic imaging techniques, and high resolution immersion ultrasonic imaging techniques. Following initial characterization; all holes were oversized to remove starter notches, the splice segments were sealed, assembled, painted, and re-characterized with immersion ultrasonic inspection. The final crack length range was found to be 0.025 to 0.375 inch.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

SAIC chose five test specimens from the 2nd layer POD study for installation of 1st layer EDM notches. These had a total of 154 fastener sites with 29 2nd layer fatigue cracks and 41 1st layer EDM notches. The target 90/95 crack size was 0.070 inch. The cracks were characterized using the same method as in the earlier study.

5.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

The test specimens used in the Battelle study consisted of sections of 10 and 12 inch austenitic stainless steel pipes. Battelle had IGSC cracks induced in the samples under laboratory conditions. Three kinds of samples were used: no cracks, short cracks 0.5-2 inch, and long cracks over 3.0 inch. A total of 80 circumferential pipe welds required 1500 operator judgments. Battelle verified crack locations prior to the study. They also verified UT characteristics to ensure that cracks, which were known to exist, were measurable using standard UT equipment and techniques. Crack locations were verified and mapped using liquid penetrant on the inner surface. Destructive testing of the specimens was planned. Battelle blocked the sample ends to prevent visual clues. The inspection area consisted of smooth surface weld and a single scribe line to be used as a reference line for measurement and recording.

5.8 MIL-STD-1823 (Draft)

MIL-STD-1823 (Draft) recommends the following guidelines.

The test specimens must reflect the structural types that the NDI process will see in application with respect to geometry, material, part processing, surface condition, and (to an extent possible) flaw characteristics. Since a single NDI process may be used on several structural types, multiple specimen sets may be required. Detailed specimen characterization is a must. Specimens should not be familiar to the inspectors, because familiarity introduces bias. Flaw sizes should be uniformly distributed on a log scale covering the expected range of increasing POD function. Very large and very small

cracks provide very little information. A specimen test set should contain at least 60 flawed sites if the system provides only Hit/Miss result and at least 40 flawed sites for a continuous response type of data acquisition. The standard also recommended that the set contain at least three times as many unflawed inspection sites as flawed sites. Unflawed inspection site need not necessarily be another specimen. If a specimen presents several locations that might contain flaws, each location may be considered a separate site.

The final geometry of the specimen shall represent the same degree of difficulty to the NDI method as the critical areas of the components to be inspected. Specimens must represent the shapes of the actual hardware for inspections where probe manipulation or inspection media are geometry dependent, (e.g., bolt holes, fillet radii, and scallops). Flaw location on specimens must be positioned and oriented to represent actual parts. The initial geometry of the specimen shall allow the insertion of flaws of required shape and size within 0.002 inch of the intended location. For UT, ET and MT methods, specimens should be from the same alloy, material form, and processing as components to be inspected. The raw material processing, heat treatment, and surface finishing for the specimen should be the same as that of the part.

Specimen handling is equally important to prevent mechanical damage and contamination. The specimens should be individually packed, carefully handled, immediately cleaned, and returned to protective enclosure after each use. The specimens should be revalidated at regular intervals. Specimen flaw responses should be measured periodically by the same test technique and procedure used for original specimen verification. The flaw response must fall within the range of responses measured in the original verification process. If it does not, then the specimen needs to be re-characterized.

5.9 Other Published Work

1988 Sample Sizes and Flaw Sizes in NDE Reliability Experiments

Sample sizes in NDI reliability experiments are driven more by economics of specimen fabrication and characterization than by the desired degree of precision in estimate of POD function [Berens 88]. Although reasonable POD function can be obtained with relatively few test results, the confidence bound calculations rely on large sample sizes. Berens emphasized that the calculation can also produce totally unacceptable results from few test samples, or from the data that are not reasonably represented by the assumptions of the models. Therefore minimal sample size requirements must be met for acceptable results. If the analysis is based on hit/miss data, then there should be a minimum of 60 flaws in the range over which the POD function is rising. Flaws outside the range do not provide as much information concerning the POD function as flaws within the range. On the other hand, if the data analysis were based on continuous response type data, Berens recommends at least 30 flaws to be present in the experiments whose results can be

recorded in this form. In practice, test sets should contain as many flawed specimens as economically feasible to increase the precision of estimates. The total specimen set should also contain twice this number of unflawed inspection sites.

1995 Sample Defect Library

Aging Aircraft NDI Center (AANC) at Sandia Labs has full-blown NDI validation equipment [Roach 95]. The FAA Sample Defect Library provides an array of full-scale, representative sections of airframe and engine structures that contain natural or engineered defects. These range from small bench-top inspectable structures to a complete B737 aircraft (38000 hours) and large fuselage sections of DC-9 aircraft. (A B747 aircraft has just become a part of this Library). These samples were acquired from various sources, and their flaw profiles were characterized for routine NDI validation process. A test specimen library database was used to manage information created and gathered by the AANC programs. AANC has structured experiments with protocols and procedures for NDI validation. These experiments assess NDI using POD and ROC, and also help investigate human factors issue and NDI reliability on different test specimens.

The Sample Defect Library was established to: (1) improve NDI development programs through the sharing of specimens and inspection information, (2) foster teamwork among FAA and aviation industry researchers, and (3) eliminate the costly production or acquisition of redundant test specimens.

5.10 Observations

- Generally, the specimens have been artificially fabricated specifically for the purpose of assessment. Some programs had a combination of synthetic and actual parts.
- Environmental degradation has generally been out of the scope of most programs. Extreme care must be taken to assure that specimen surface condition is not altered during an assessment program and to provide rigorous cleaning to removed residual penetrant materials from cracks between inspection sequences.
- Different programs have varied substantially in the number of specimens used. As the statistical tools evolved, the specimen requirement reduced.
- The concept of racks with interchangeable subassemblies is a good way to eliminate use of prior knowledge from inspectors who have already gone through the program.
- MIL-STD-1823 (draft) makes valuable recommendations for specimen configuration and handling. They are still good.
- Specimens must be carefully characterized and documented before use in an inspection program. The specimens and the related documentation need to be preserved in a centralized Library form for future use.

Best Recommendations: MIL-STD-1823 (draft) or MIL-HDBK-1823

6. INSPECTION SCHEDULING

Statistically significant number of specimens and inspections are required for generation of a POD curve at a 95% confidence level. Each inspection takes a finite amount of time. If multiple facilities need to be visited, the inspection duration and scheduling impacts the total program cost and time-span. Reliability assessment programs are also known to disrupt the daily production line inspections. Scheduling is therefore an optimization problem.

6.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

The entire Lockheed program lasted from June 1975 to January 1978. It began with a "dry run" on June 16-26, 1975, at Wright-Patterson Air Force Base and continued with visits to 21 other facilities covering all five AFLCs, four ATC bases, four MAC bases, four SAC bases, and four TAC bases. Almost 300 NDI technicians completed approximately 800 separate inspection tasks and over half a million inspection sites.

The typical stop at each location was 4 weeks (20 working days). During the 20 working days, the inspection schedules allowed parallel inspections. Rotation from one specimen to another optimized the scheduling efficiency. The structural samples were transported to the Air Force installations in a 16 foot utility trailer towed with a pickup truck.

1979-84 NDI Technician Proficiency Program

Lockheed planned the proficiency tests to be completed in a normal working day with normal breaks. The plan provided a minimum of 7 hours for each test set. No test was allowed to extend into the second day.

1987-88 Engineering Services in Support of NDI

Lockheed imposed no time constraint when conducting the total exercise. In a normal evaluation, the individuals were allowed a maximum of 4 hours for radiographic and eddy current NDI for a fastened joint standard, 4 hours for eddy current NDI for a lug standard, 4 hours for penetrant NDI for a flat plate standard, 6 hours for magnetic particle NDI for a lug standard, 8 hours for ultrasonic NDI for a honeycomb standard, 8 hours for a fastened transition joint, and 8 hours for a bonded transition joint standard.

6.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

The total Martin Marietta program ran from Sept 1978 to June 1982. The actual on-site assessment data collection took two months each at SA-ALC and OC-ALC during the summer of 1981. The program team planned each inspection cycle, running over 200 inspection opportunities, for a 4-hour sequence.

6.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

During Battelle's initial visits to the AF field labs, discussions indicated that a 1-day disruption in the work load was moderately acceptable to the lab supervisors, but 2 or 3 days would be considered more than what could be accommodated easily at many labs. Battelle then considered two types of work schedules: a nominal schedule to be done within an 8-hour day and ambitious schedule up to a 10-hour day running through the coffee breaks and lunch. The proposed schedule was based on one practitioner per method per hour, on an average. The following four programs were proposed: (1) 6 hours, five NDI methods and five technicians, (2) 6 hours, three NDI methods and six technicians, (3) 10 hours, five NDI methods and eight technicians, and (4) 10 hours, three NDI methods and nine technicians.

6.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

Not applicable.

1986-88 NDI Personnel Proficiency Evaluation Using UT

The program gave each technician sufficient time to complete an individual test. However, total time was limited to 8 hours in phase I and 4 hours in phase II.

1990-94 Reliability Assessment Kit

Not applicable.

6.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

The program for FAA included visits to nine facilities between April 01, 1993 through August 13, 1993 to conduct ECIRE. The visit was approximately 1 week at each facility. The tasks were designed to take at least an estimated 4 hours to complete; however, the inspectors were given as much time as was needed to complete the task. To understand the impact of shift work on reliability, the program team attempted to perform inspections in the graveyard shift wherever possible.

6.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

SAIC made no mention on scheduling or time limitation on inspection.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

SAIC made no mention on scheduling or time limitation on inspection.

6.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

During the round robin, Battelle encouraged the technicians to complete each inspection within 45 minutes; however, they did have as much time as they desired to make an inspection.

6.8 MIL-STD-1823 (Draft)

No inspection scheduling constraints are mentioned in this document.

6.9 Other Published Work

1999 An Interview with an Ex-USAF Technician

The main concern expressed by one of the AF personnel who had traveled with specimens to various facilities for ET POD data generation was that at most facilities, inspectors had no spare time at all for the program. Though they were very keen on

knowing how well they were doing, the normal work load gave very little room for participation in such a program.

6.10 Observations

- Most of the programs provided ample time for inspectors to complete their inspections within a realistic upper bound.
- The general constraint appears to be inspector availability as offered by their supervisors, and without significantly disturbing the production line.
- In production environments, accelerated inspections are performed for various reasons on many occasions. This aspect does not appear to have been addressed in any of the programs.
- Impact of shift time was studied by FAA and Martin Marietta.
- Impact of overload was studied by Battelle.
- Scheduling should address all issues that are local to the facility such as overtime, shift work, and accelerated inspections.

Good Practice: FAA Program

7. INSPECTIONS

Inspections are the execution phase of reliability experiments. A well-planned experiment with well-designed specimens must be followed by equally well-executed experimental efforts.

7.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

The inspection techniques employed for the Lockheed program were UT, ET, PT and MT, but all methods were not applied to all specimens. Lockheed based their selection on actual field practice.

Lockheed provided the structural samples to NDI technicians in settings that very closely represented those encountered in routine field and depot installations. The team placed some samples in an overhead position to simulate NDI on a wing lower surface. Other configurations included face-up positions and vertical plane positions, typical of a full range of structure. A validation run on each procedure was conducted at Lockheed to ensure compatibility with the program objectives.

The team gave all participating NDI technicians the same program orientation using an audio-video system (a combination of 35-mm slide and a tape playback). They assigned specific NDI tasks on the samples after the orientation briefing. The program team also formulated complete NDI procedures in the T.O. format per MIL-M-38780A for each NDI method as applied to each structure type. The procedures included necessary operating parameters and equipment calibration details. Participants had ample opportunity to read instructions and ask questions during the briefing.

At each installation, participants used existing equipment. This helped provide an indication of equipment condition on NDI reliability.

The participants performed inspections under the guidance and coordination of the accompanying Lockheed engineer. However, the goal was to avoid assistance that could bias the results.

The technicians marked the detected cracks with grease pencils. These marks were removed by a solvent to prepare for the next technician. The residual penetrant from specimens was removed using an isopropyl alcohol bath. Lockheed verified that the repeated cleaning did not introduce any biasing.

During the program, Lockheed added certain newly developed NDI equipment and specially fabricated test specimens. They added automatic eddy current bolt hole inspection in December 1976 so that comparisons could be made between newer automatic techniques and the standard practices. They added seventy six technician

proficiency-screening samples in January 1977 to determine if a relatively simple, small inspection sample could be used to evaluate or predict technician NDI ability on larger, more complex structure.

At the end of the inspection tour, Lockheed subjected the samples to a tear down inspection to confirm flaw locations and identify flaws at possible sites that had a high rate of "false calls."

1979-84 NDI Technician Proficiency Program

For this program, Lockheed initially briefed technicians to make them realize the need and importance of measuring NDI capability on an individual basis. The briefing then provided the instructions for technician participation in the demonstration of capabilities, the methods for determination of results, and the design of a uniform qualification program based on the results. It also covered the detailed NDI procedures.

The tests involved activities that were normally encountered in the NDI process: reading and understanding the procedure, setting up the equipment and calibration, discriminating between a true flaw and a false call, and reporting results. Lockheed took into account the effect of false calls in scoring the results.

The individuals taking the test identified the flaw locations and graphically recorded those findings on a data sheet (schematic drawing of the test object). The test hardware needed to be inspected in its normal, assembled upright position.

The technicians inspected the laboratory fatigue cracked specimens. The NDI procedures applied were not typical of routine inspection, and the technicians performed solo without any backup confirmation of results.

1987-88 Engineering Services in Support of NDI

Lockheed required for this program that all inspections be performed in locations qualified for the task, with adequate space available. If that involved performance in the NDI shop area, then special care was required to maintain separation of the evaluation process from other activities. The details of the various NDI procedures are clearly defined in the report [Sproat 88]. The program administrator monitored the activity as a bystander. He assembled the hardware, inspected specimens for any marks that could be construed as flaw location indicators, inspected the exercise, and assured the specimens were left clean. Lockheed maintained short-term storage in an area free of high humidity, potential for damage, and presence of any corrosive media. They maintained long-term storage by re-packing as received and placement in a controlled environment.

7.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

This Martin Marietta program limited its scope to initial tear-down inspection operations for components in the gas path. Inspection methods assessed included fluorescent penetrant[5] (FPI), ET, UT, and MT. The program team gave special attention to FPI, since the majority of inspection operations in engine maintenance facilities use FPI. The program did not include X-ray, since laboratory personnel performed X-ray off-line.

Communication was recognized as a critical factor for success of the project and timely utilization of the information gathered. Extensive briefing sessions were a part of the program. The team distributed a preliminary briefing document ahead of time to familiarize management with goals of the program. They also presented on-site briefings for management at the two ALCs and the AFLC headquarters. Automated slide-sound presentations formed the operator briefing program.

The program included a random specimen selection from test sets to provide approximately 100 flaws and 200 inspection opportunities.

Each test set was processed as uniformly as possible in the production-line facilities utilizing the inspection materials, processes, and operating procedures that were in general usage. Each test set was tracked through all process steps to assess the capabilities of individual processing steps and the total process, to observe consistency of variables in process application, and to reduce the risk of the loss of a test specimen in processing lines. Equipment performance assessment and measurement were completed and documented for each processing station.

The FPI was applied to most engine hardware and thus constituted the largest number of assessment cycles and data. Other processes used were ultrasonic surface wave, MT, and ET.

While performing the reliability assessment, directly correctable variations in performance were reported to the primary facility contact at each facility as the variations were noted. Data were collected after such corrections were made. Direct observations made during data collection were presented during exit briefings at each facility.

Management at each facility was briefed on the direct observations and preliminary assessments made at each facility at the conclusion of facility data acquisition phase. Results of preliminary analyses, direct observations, conclusions, and recommendations were presented during a requested preliminary briefing at both facilities and at AFLC headquarters at the end of the program.

[5] Same as PT

7.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

Battelle surveyed and reviewed the AF personnel in their process assurance program. The observations were performed at AF labs and field trials, as discussed below.

Observations at the first six labs: At each AF lab, the Battelle visit started with an introductory discussion with the lab supervisor or his designee. The discussion began with review of AF Program Office's overall NDI process assurance program and concluded with the specific goals for NDI process assurance for field labs. The purpose of these visits was to gather information and exchange ideas about process assurance. The lab supervisor then reviewed the NDI activities, facilities, equipment, and staff. A tour of the lab and observation of the ongoing activities followed, including a visit to a maintenance hanger. During these tours most of the on-duty airmen were introduced and encouraged to participate in discussions about their NDI activities and training. Various trial concepts involving NDI process assurance were presented and spontaneous reactions to proposed concepts were actively solicited. Each visit lasted 4-5 hours on same day.

Field Trials: Upon arrival, the Battelle team briefed the lab supervisor on the background of the NDI process assurance program, goals for the visit, and the nature and scope of the planned activities. The team assured lab personnel that the purpose of the visit was to resolve the logistic and procedural details and not to measure the individuals or the lab. This took about 15-20 minutes and was followed by a general briefing to the on-duty NDI practitioners. Although they were asked to put their name and base on each of the data sheets, identification was to facilitate record keeping and was not to be disclosed. The team requested the participants not to discuss the nature or results of any particular sample set until all participants had examined it. This request was followed by a more detailed discussion about the samples, general design of samples, general procedures to be used in examining the samples, and use of data entry sheets for each sample type. The team emphasized the use of everyday NDI techniques for examining the samples and asked the participants to quantify to the extent possible the size and location of any defect indication that they marked on their data sheets. Each participant was encouraged to select a sample set with which to start. During the inspection procedure, the Battelle team remained in the lab, observed the ongoing activities, answered questions, and as necessary provided additional guidance and advice related to the examination of a particular sample. When finished, each participant returned the data sheets to the investigator and moved on to the next set.

A participant’s proficiency had significant impact on the time taken to complete the test. The test times varied between 45-90 minutes, with an average close to a pre-trial estimate of an hour. Handling of the samples was the greatest difficulty observed. Apparently, most of the participant's experience had been with large monolithic and discrete items; in this instance, the test pieces were several small coupon samples. Participants also had difficulty discriminating between a spurious and true indication.

7.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

Not applicable.

1986-88 NDI Personnel Proficiency Evaluation Using UT

SwRI gave individual NDT technicians both written and verbal instructions on how the proficiency evaluation was to be administered. The inspectors were to perform A-scan UT using a surface wave technique in accordance with a T.O. 2J-F100-9. The inspection methodology was similar to what the inspectors normally used in the production environment. The T.O. called for UT instrument calibration to be performed on a test block with a 0.050 inch Elox slot, and the amplitude of the instrument on this flaw was set to 50%. The technicians were told to suspect signals above the limit, but to reject any blade with a 75% amplitude or greater.

1990-94 Reliability Assessment Kit

The SwRI Team prepared procedures for each inspection method using MIL-M-38780 as a model, and writing them in the form of a T.O. Details of the T.O.s are given in the appendices of the report [Goodlin 94]. SwRI validated each procedure at SwRI using a certified level III inspector who developed the procedure. Validation was performed to verify the location and size of programmed flaws in the standards. The validation was conducted as much like an actual examination as possible. SA-ALC also furnished personnel to setup and administer the validation.

7.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

Two monitors traveled with the experiment to all nine facilities where the ECIRE was to be performed. The experiment allowed inspections to occur as they normally would. Therefore the equipment and procedures followed were of the inspectors own choosing. The only potential departure from routine inspection was in the way they recorded the results of their inspection.

To address the issue of reliability and uncomfortable posture, half of the skin specimens were presented at 24 inch height (knee level). To investigate repeatability under identical conditions, the first inspector at each facility was returned to repeat the experiment after all four of the initial inspections were complete. The order of skin specimens was altered between inspections. The monitors promised complete protection to an inspectors, and their performance was not reported to the facilities.

7.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

SAIC automated the splice joint inspection process to the greatest extent possible. All inspectors were required to use C-141 lower inner wing spanwise splice joint second layer fatigue crack ultrasonic inspection procedure. All inspectors used the same calibration standards with EDM notches. All inspections followed the protocol developed from FAA's generic protocol [Spencer 93]. The inspections took place on crack specimens mounted overhead to simulate a C-141 lower inner wing. Actual full length panels were mounted on a framework to simulate work on typical C-141 wing stand platform. The specimen placement could be varied along the wing panel length. The inspector mounted the scanner on the wing panel and inspected the crack specimens in the same manner as would have been accomplished on an aircraft.

The laboratory validation was intended to characterize the impact of procedural variables on detection as well as quality of signal. Fractional factorial experiment was designed to cover five procedural variables, which were time-base delay, depth velocity, receiver gain, scanner skew, and probe pressure. Seventeen experimental runs were made with two sets of specimens. Each run was a scan of all the fastener sites included in specimen subset.

In the field portion of the validation, individual inspectors performed a complete inspection of the test specimens following the procedures that would be used in an aircraft inspection. Eight specimens were butted and hung from the frame to simulate the lower surface of the wing. In all cases, inspectors had no prior knowledge of the crack distribution. In addition to making calls on the data gathered during their own inspection, 10 of the 16 inspectors were asked to bring up a set of images and make calls on them. By including a common set of data, the decision component of the inspection could be separated.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

The procedure was a repeat of the previous program.

7.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

Battelle presented the pipe specimens to technicians in a random sequence, on demand. The technicians conducted the inspections on a work bench, with freedom to roll over the specimen. Each technician was requested to bring his own equipment. He could also use the Pacific Northwest Laboratory (PNL) equipment. The bench space was adequate to place both his and PNL equipment. Space was also available for calibration blocks, and

frequent calibration checks were common. Lighting and temperature were at a comfortable level in the laboratory. The inspection workstations were set up to remove as many extraneous influences as possible from the detection task.

After the inspections each technician was engaged in a one-hour interview with a human factor specialist and asked to fill out a survey form containing questions on their background and experience.

7.8 MIL-STD-1823 (Draft)

MIL-STD-1823 (Draft) recommends:

The procedures to be used in a demonstration must follow the procedures and work instructions planned for the production inspection of parts. This requirement includes all fixed process parameters, data analysis algorithms, accept/reject criteria, and any other items covered by a control document. The inspections shall be performed by production inspectors. A test monitor shall be designated who shall assure that all requirements of the MIL-STD are being met both prior to initiation and during the performance of the tests. Specimen fixturing and actual component should have the same inspection system arrangement of probe orientation, manipulation, and scan plan. The data should be recorded in a form that is compatible with the disposition of the part. There should be a plan to ensure state of control throughout the demonstration experiments. The plan should include routine quality, instruments, and calibration checks and should also include inspection responses to real structures. The process control plan should be the basis for process control during extended periods of production inspections.

The sets of inspections as defined in the demonstration design document shall be carried out at the production inspection facility under normal operational conditions. A test monitor shall be available during all testing. A log should be kept of the inspections showing the order in which the inspections were performed, the inspector, the specimen identification and serial number, and the date and time of inspection.

7.9 Other Published Work

Nothing relevant was found.

7.10 Observations

- Most of the programs involved testing in an actual work environment with actual equipment that was used in production lines. These criteria are truly important in order to obtain a realistic reliability estimate. Briefings to management and participating inspectors are an important part of the actual inspection part of the program.

- If extensive details on how to inspect are provided and total compliance is ensured, performance better that normal will usually be obtained. For more truthful assessment, inspectors should be allowed to follow their routine process and procedure.
- Isolation of results from one inspection to another is very important to get independent data points. This requirement necessitates thorough cleaning of specimens to remove any markings that can bias the results.
- Inspectors need to be assured that their identity will not be disclosed and their performance will only lead to a data point in the analysis. Still, human factor specialists say that an inspector's test performance deviates from his routine performance.
- Most programs had an onsite monitor to answer queries from the participants. However, at times these monitors may have had a tendency to bias the results.
- Specimen handling from unpacking through setting up, cleaning, and re-packing needs special attention to avoid any changes in characteristics.
- Observation of facility characteristics is an important segment.
- Assessment of inspection materials, calibration artifacts, and inspection equipment is required to identify variant data.

Best Practice : FAA Program

8. DATA ACQUISITION AND HANDLING

The value of the experiment is in the data obtained from the execution. The constituents of data acquisition are (1) the type of information recorded such as inspection results, controllable and uncontrollable factors within or outside the experiment, and (2) the form of information records such as graphic or textual.

8.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

In this program, technicians at AF bases examined samples for cracks and marked them with grease pencils. After each base visit, the base transmitted the raw data from the field to Lockheed by mail or by a carrier. Lockheed systemized the flaw indications for data processing in a standard format. Technician profiles, equipment evaluations, and environment reports were filled by AF categories in their original narrative form.

The AF Logistics Command Program Monitor archived all raw data classified by location. Also the raw data are maintained.

The nucleus of all data acquired was composed of "Find" or "No Find" information for each catalogued flaw site inspected by each participant. Flaw sites catalogued at the start of the data acquisition phase were those identified as suspect fatigue cracks of lengths estimated initially by NDI methods. Subsequent structure tear down and detailed examination at the conclusion of the data acquisition phase provided more accurate data on flaw content.

A general purpose *System 2000* Data Management System, developed by MRI Systems Inc., Austin, TX. was used for data handling. The system capabilities included storage and organization, updating with new inputs, identification and isolation of important data, simple mathematical and statistical computations, and report generation for decision making. Most of the internal programming was transparent to the user. As this project was performed in 70s, the system must have been the best available at that time. Details however, will not be discussed further, since present-day computers have far more potential and the actual techniques employed then have very little significance today.

The data acquired were classified in seven categories: (1) flaw-size tabulation; (2) inspection results - finds, no finds, and false calls; (3) individual inspection log; (4) technician profile; (5) base daily; (6) facility evaluation; and (7) equipment performance.

1979-84 NDI Technician Proficiency Program

During the test, four forms were filled out:

Administrator's Log: Record of daily activity, participant identity, NDI method used, test configuration, and total time spent on the inspection task.

Participant's Log: Record of actual test time.

Equipment Log: Record of test equipment.

Data Sheet: Schematic drawing of the test object to mark sites of flaws.

The individuals taking the test graphically recorded their findings on a data sheet (schematic drawing of the test object). They were not required to indicate length or direction of flaws, only the site of detection. Grading was done in terms of finds, misses, and false calls. The performance was measured by a coefficient of contingency, which exacts a penalty for false calls. No flaw length factor was attached to the performance determination.

1987-88 Engineering Services in Support of NDI

The participants were to indicate flaw "finds" on a flaw data reporting sheet, one for each rack of specimens. Different sheets for different specimens had the specimens sketched in a way making it very convenient for the inspector to mark identified damages on them.

8.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Martin Marietta used a portable microcomputer data-entry stations for data collection and documentation. All inspection data were entered into a single format, matrix-type database system. Data entry was prompted using graphics and menu selection (advanced at that time). The data programs and data storage were compatible with most computer systems that supported BASIC language.

A brief operator experience profile was recorded for each operator; however, care was taken to prevent any association of inspector identity with the data.

After over 2 months each at Tinker AFB and Kelly AFB, the team completed and documented for assessment 171 total test sequences by 113 different inspectors. Data constituted approximately 35,000 inspection opportunities for flaw detection.

All test sets had been characterized and flaw sizes documented prior to their use at AF facilities. Re-verification was performed by surface replication to document and confirm actual crack sizes after completion of the inspection cycles. Fluorescent penetrant was used to confirm that cracks were open and help in surface replication.

The team measured and tabulated crack length data and used the data as actual crack-length dimensions for all assessments.

8.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

Battelle had inspectors enter the crack data onto sheets provided along with the specimens. Sheets had graphic pictures of the specimens. Battelle encouraged participants to indicate quantitative measure of the crack on the sheet wherever possible.

8.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

Not applicable.

1986-88 NDI Personnel Proficiency Evaluation Using UT

Data were collected in form of finds, misses, false calls, and correctly identified unflawed sites.

1990-94 Reliability Assessment Kit

SwRI developed computerized data-entry sheets for each test specimen for on-site use. Technician number, specimen number, and flaw location could be entered into the sheet. Simple checks in the algorithm allowed for quick identification of coding errors. The software included modules that guided the data-entry operator through the inspection report form for each combination of specimen and inspection technique. SwRI designed the output to be ready for the data analysis algorithm.

8.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

For the ECIRE project, inspectors marked directly on a piece of protective tape (Scotch™ # 336) that the experimental monitors put into place before each inspection. The tape protected any visual clue left on the bare aluminum during pencil marking. The tape had virtually no influence on the magnetic fields. The inspectors were also asked to give a subjective rating (1, 2, or 3) reflecting their confidence that a flaw signal was present.

In addition, monitors recorded supporting data that did not form part of the controlled factors. These supporting data were factors that could not be controlled in the experiment such as environmental factors like humidity, temperature, noise, light, and housekeeping and personnel factors like age, sleep, gender, posture, and psychological. These factors

were limited to those that could be easily measured and recorded or obtained through a self-report from the inspector.

8.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

A Pentium processor was used for automatic data acquisition.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

Same as previous.

8.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

For the UT round robin, Battelle asked each technician to record his own data. The data were entered on a company form and was later transcribed by the technician on the PNL form for evaluation. In addition to recording information on crack location and depth, each technician was required to indicate the degree of certainty of crack detection.

8.8 MIL-STD-1823 (Draft)

MIL-STD-1823 (Draft) recommends:

The inspector should prepare a report on each inspection performed. The report should contain the inspector identification (possibly coded), specimen identification, inspection date and time, and results of the inspection including the NDI responses and locations of any indicated defects.

8.9 Other Published Work

Nothing relevant was found.

8.10 Observations

- Different programs had different ways to record defect data.

- Initial programs only concentrated on finds and misses, but later efforts recorded false calls and truly identified flawless specimens, as well as their level of confidence in the inspection result.
- Graphical form for defect data recording is probably the best; however that may deviate from the actual practice that an inspector might normally follow.
- The physical environment such as temperature, humidity, time of test needs to be recorded in some sort of a form so as not to miss important data bearing on the inspection.
- If possible, inspectors should be asked to fill out a questionnaire to record their physical and mental state at the time of the test.
- All tests must record all four bits of information - finds, misses, false calls, and true no-calls.
- Flaw sizes for finds and false calls should be recorded wherever possible (for $\hat{a}$ versus a type of analysis discussed in the next chapter). Since this is generally not a normal part of most operations and it may perturb the normal inspection operation.
- All data should be archived and stored for subsequent analysis in a media that can last for many years to come (hard copy perhaps).

Best Practices: SwRI for data acquisition. FAA for type of data recording

9. DATA ANALYSIS

NDI reliability experiments can be categorized in three types: (1) demonstration of capability at one crack length, (2) determination of POD function through single inspection of cracks covering a range of lengths, and (3) estimation of POD function and confidence bounds through multiple inspections of cracks covering a range of lengths. Analyses of data from category (1) and (2) are generally based on the binomial distribution theory, which groups cracks within an interval. Category (3) data are analyzed by fitting an equation to the observed detection probabilities for each of the cracks (regression method). The latter procedure considers the scatter of detection probabilities at each crack length caused by the non-reproducibility of all factors other than crack length [Berens 83].

Generating a POD curve involves two basic steps: (1) Generate a point estimate of detection probability for various discrete crack lengths over the range of interest, and then (2) Fit an appropriate curve (functional form) that offers minimum deviation or maximum likelihood to the scattered data. Most of the research over the years has been in arriving at the best functional form that can help generate the POD curve with an affordable number of experimental data points. The group at UDRI extensively studied the subject and developed mathematical methods [Berens 81, Berens 83, Berens 84, Berens 84a, Berens 88, Hovey 88, Berens 97] as well as computational tools [Berens 88a] to perform data analysis. Two techniques became acceptable for estimating POD as a function of crack size. In the past, the techniques available for estimating the false call rate have been geared more towards operator performance than evaluation of the NDI system reliability.

9.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

A convenient format for performance evaluation was either a curve or a histogram of flaw detection probability relative to flaw size (flaw length in this case). In the beginning, data on detected flaws were compiled in the form of composite histograms encompassing an average of seven individual flaw lengths. Seven point grouping was selected arbitrarily from a viewpoint of convenience. As data acquisition progressed, the number of individual attempts to find each flaw was sufficient to develop analytical approximations of detection probabilities relative to each of the flaw lengths. This provided the individual observed point values of POD for crack lengths, with all the scatter inherent to any POD study. Lewis et al. examined a number of transformations of the variables to select those that minimize scatter about an estimated mean. They

observed a reasonably good analytical fit with the following functional form and transformations:

$$\text{POD}(a_c) = e^{-a a_c^{1-b}} \quad ; \qquad y = \ln\left(\frac{-\ln \hat{p}}{a_c}\right) \quad ; \qquad x = -\ln(a_c)$$

where a_c is the flaw-size parameter and $\hat{p}$ is the point estimate of the detection probability for crack length a_c. Constants α and β are determined using linear regression analysis of the data. Further details and statistical measures of the scatter are mentioned in the full report [Lewis 78].

1979-84 NDI Technician Proficiency Program

Lockheed entered the results from inspections in terms of finds, misses, false calls, and correctly identified non-flawed sites into a statistical formula that provides an index or factor of relative performance. This factor of performance provides a uniform system of grading and a method of comparison of technician capability within the AF. Once sufficient data are collected, performance levels can be established and results can be applied to identification of training requirements.

Chi-square form of data analysis accounts for false calls. Lockheed accomplished this method by comparison of performance with what could be expected by "lucky finds" attributed to chance. The analysis begins with a decision matrix for the test data as shown.

	Flawed	Unflawed	
Marked	Find (A)	False call (B)	R_1=A+B
Unmarked	Miss (C)	Correct no-call (D)	R_2=C+D
	C_1=A+C	C_2=B+D	N=A+B+C+D

The key matrix entries are the four combinations of flawed/unflawed hardware and marked/unmarked NDI technician response. Coefficient of contingency is then defined using the chi-square value to estimate the contribution of chance, as

$$\text{Coeff of contingency} = \sqrt{\frac{1}{C_1 R_1}\left(A - \frac{C_1 R_1}{N}\right)^2 + \frac{1}{C_2 R_1}\left(B - \frac{C_2 R_1}{N}\right)^2 + \frac{1}{C_1 R_2}\left(C - \frac{C_1 R_2}{N}\right)^2 + \frac{1}{C_2 R_2}\left(D - \frac{C_2 R_2}{N}\right)^2}$$

This coefficient is an index of performance and varies from 0 to 1. The lower the value of the coefficient, closer is the performance to chance. This method leads to a lower rank for the technician who may exhibit high detection probabilities at the cost of false calls. As mentioned earlier, this mathematical formula is geared more towards technician performance than evaluation of the NDI system reliability.

1987-88 Engineering Services in Support of NDI

The report only says, "For evaluation purposes, finds, misses, false calls, and correctly unmarked/non-flawed locations are statistically treated to establish performance levels." And it also says that "the basic measure of NDI performance with fatigue-cracked standards is the ratio of finds to total flaws."[6]

9.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Graphical data output from application of an inspection procedure provided a quantitative measure of the performance of NDI procedure as a function of flaw size. Martin Marietta project team obtained a single data point on the POD curve by starting at the largest actual flaw size in a group, counting down 29 detection opportunities, calculating the point estimate of detection (detected/opportunities), and plotting the point estimate of detection for the POD curve. They plotted the POD curve by curve fitting to the series of point estimates. Then they applied various curve-plotting routines to smooth the data for both individual and combined curves. Curve-fitting routines evaluated included linear regression analysis, high-order polynomial fit, a Lockheed model (discussed in the previous section), and a log-odds model by Berens and Hovey (to be discussed later in this chapter). These methods were shown to be crack distribution dependent and therefore did not qualify as general models. Segmented versions of the Polyfit, Lockheed, and Log-odds methods decreased the effects of the various crack distributions and provided a better data fit. The team then plotted the actual data and presented in the report along with the modified least square curve fit. They determined the threshold detection level to be the flaw length corresponding to the inflexion point on POD curve.

Performance capability and reliability were initially obtained by plotting POD curves for each inspection sequence. Comparison of POD curves for different inspection sequences performed on a single test set at a single test station provided both qualitative and quantitative measures of process variations due to human-factor variables. Such comparisons are valid only for controlled processing sequences. Some variability could be attributed to human-factor variables for some inspection sequences, and such variability was usually accompanied by a high false call rate. Measured crack data from single inspection sequences were combined for similar inspection types to obtain composite POD curves for overall detection performance. The combined curves reflect a measurement of overall capability, but are of a limited value for analysis if data from uncontrolled and controlled processes were combined. Combined curves, however, could provide a quantitative measure of process improvements when compared to a previous combined data curve (baseline).

---

[6] Further details on data analysis are probably in [Sproat 88b], which could not be accessed at this time

9.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

Battelle developed a software package using Lotus 1-2-3 spreadsheets to run on IBM PC. Maximum likelihood POD estimates, ROC curves, and similar data analysis functions could be performed with this software. In addition, chi-squared values were generated and used to indicate 90% level of confidence. With the limited data set, Battelle found some of the outliers had adverse impact on POD estimates.

During the later part of the work (1988-89), Battelle made efforts to improve the procedure for estimating POD curves [Hyatt 89]. Two outlier-resistant estimation procedures performed satisfactorily on the given data set, though they were not fail-safe. The report recommended that using better physical models of the inspection process, increasing the size of data sets, and using better choices of crack lengths in the tests would provide additional resistance to the outliers.

Hyatt investigated [Hyatt 91] three procedures and reported that two of them can provide POD estimates with small data sets that contained rogue points, unlike the Maximum likelihood method of Berens and Hovey [Berens 84]. The latter requires large data sets to produce reliable POD curves. The procedures essentially consist of attaching a weighted function wherein rogue points get identified mathematically and collect higher weights, thus having little effect on resulting estimates. These procedures are named the Huber and Cosine Taper methods. The mathematical description is too complex to be explained in brief in this section. The power of the methods described was demonstrated using data collected from field tests of aircraft engine line inspectors applying fluorescent dye penetrant technique.

9.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

There were no recommendations on data analysis procedures.

1986-88 NDI Personnel Proficiency Evaluation Using UT

Two different types of analyses were performed: (1) the coefficient of contingency as described by Lockheed and (2) the ROC curve with limits of proficiency at 80% POD and 30% POFA. These analyses were done for the full specimen set in Phase I, specimen subset in Phase I, and full specimen set in Phase II.

1990-94 Reliability Assessment Kit

SwRI developed data analysis algorithm to provide graphical output illustrating the estimated POD of a given flaw size. They performed analysis using all inspection data received from a test site for each of the NDI test kits. The method of estimating the POD function from the inspection data is based on procedures outlined in the report "Flaw detection reliability criterion" [Berens 84]. The reliability data collected on the inspected cracks are in pass/fail format with multiple inspections for each crack. Both maximum likelihood and regression analysis techniques can be used to estimate the POD function when each flaw has multiple inspections. However since maximum likelihood estimates are difficult to calculate when only a few inspections per crack are recorded, only regression techniques are used to estimate the POD. The details of the log-odds method are provided later in this chapter.

9.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

Binary regression models for hit/miss data are written in the form

$$\text{PoD}(a) = \Pr(Y = 1 \mid a) = F(a + b.\log(a)); \quad \text{where } Y = 1 \text{ for hit and } Y = 0 \text{ for a miss}$$

Where a and b are parameters to be fitted to the data and F is the cumulative distribution function. The choices for the distribution function F were:

$$\text{Normal } \int_{-\infty}^{x} \frac{e^{-z^{2/2}}}{\sqrt{2p}} dz; \qquad \text{Logistic } \frac{1}{1+e^{-x}}; \qquad \text{Gompertz } 1-e^{-e^{x}}$$

These models are sometimes developed in terms of Log(a) rather than 'a'. They are then referred to as lognormal, log-logistic, and Weibull forms. According to Berens and Hovey [Berens 83], log logistic was the best model among seven different functional forms for POD tested for their NDE application.

The basic POD model could be extended to include the explanatory factors by adding more parameters that denote the state of other factors present at the time of inspection. The concept comes from the motivation to be able to alter a and b according to different conditions under which inspections occur. Such an alteration, however, would require adequate data. ROC curves were generated to account for false calls. Background discussions on POD curves can be found in [Berens 88, Hovey 88, and Annis 89].

9.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

The model used for maximum likelihood estimation was a four-parameter adaptation of the probit model given by

$$POD(a) = a + (b - a).\Phi(c + d.\ln(a))$$

where Φ is the standard normal distribution function; a is the crack length; and α, β, c, and d are the parameters to be fit. The other constraint is $0 \le a \le b \le 1$. The parameter α is generally interpreted as a false call rate.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

Same as previous.

9.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

The assessment divided specimens into 1 inch segments; each segment organized the data in the form of a hit, miss, false call, or true negative. ROC analysis was chosen over the more traditional approaches for measuring NDT technician performance accuracy because (1) traditional methods are unable to account for different flaw densities, (2) accuracy scores do not reflect different patterns of correct and incorrect decisions that can occur, and (3) traditional methods do not distinguish between technician's ability to discriminate crack-induced signals from non-crack-induced signals. Two separate indices of performance were computed in each ROC curve: (1) the area under the ROC curve and (2) the probability of a true crack call when the probability of a false call is 0.20. The area under the ROC curve = 1 represents perfect performance and 0.5 represents a pure chance. The second index is of interest, since the passing grade for performance demonstration tests requires that the false call rate be 20% or less and the hit rate be 80% or more. For one-point ROC measurement, the second index is a better representation. Swets presents a complete description of ROC analysis in the context of UT [Swets 83].

9.8 MIL-STD-1823 (Draft)

MIL-STD-1823 (Draft) recommends that the analysis of the collected data should be accomplished using a standard IBM PC program supplied by USAF (ASC/ENFP WP AFB) and named POD/SS. The program will not work if data are missing. A statistician may be required in such a situation.

9.9 Other Published Work

1981 Statistical Methods for POD Estimations

Berens and Hovey presented a statistical framework for describing the uncertainty in NDE determinations and evaluated various characterizations of NDE reliability [Berens 81, Berens 83]. The data from "Have Cracks" program were used to estimate the parameters of the NDE model. For the representative capabilities, NDE reliability experiments were simulated. Different NDE capability characterizations were computed for each simulated experiment and were statistically compared. The paper demonstrates that the regression model is appropriate for characterizing the POD function.

Berens and Hovey [Berens 81] investigated seven formulations with functions and corresponding transformations. Regression analyses were used to fit all seven models to the "Have Cracks" data. The mathematical formulations of these models were

Lockheed $\mathrm{POD}(a_c) = e^{-a a_c^{1-b}}$; $y = \ln\left(\frac{-\ln \hat{p}}{a_c}\right)$; $x = -\ln(a_c)$

Weibull $\mathrm{POD}(a_c) = 1 - e^{-a a_c^{b}}$; $y = \ln(-\ln(1-\hat{p}))$; $x = \ln(a_c)$

Probit $\mathrm{POD}(a_c) = \Phi(a + b a_c)$; $y = \mathrm{Probit}(\hat{p})$; $x = a_c$; $\Phi(x) = \int_{\infty}^{x} \frac{1}{\sqrt{2p}} e^{-1/2 y^2}\, dy$

Log Probit $\mathrm{POD}(a_c) = \Phi(a + b \ln(a_c))$; $y = \mathrm{Probit}(\hat{p})$; $x = \ln(a_c)$

Log-odds(linear) $\mathrm{POD}(a_c) = \frac{e^{(a+ba_c)}}{1+e^{(a+ba_c)}}$; $y = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$; $x = a_c$

Log-odds(log) $\mathrm{POD}(a_c) = \frac{a a_c^{b}}{1+a a_c^{b}}$; $y = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$; $x = \ln(a_c)$

Arcsine $\mathrm{POD}(a_c) = \sin^2(a + b a_c)$; $y = \arcsin\left(\sqrt{p}\right)$; $x = a_c$; $0 \le a \le \frac{p - 2a}{b}$

The detection probabilities $\hat{p}_i$ and the crack lengths a_{c_i} for each crack were transformed to y_i and x_i in accordance with the transformations. The transformed *x* and *y* variables were then used in a linear regression analysis of the form $y_i = A + Bx_i + e_i$. For all seven models, B is the estimate of β and, depending on the model, either A or e^A is the estimate of α. The deviations e_i, of the transformed observations from the regression equation were analyzed to test the applicability of each model with respect to a pre-defined

acceptability criteria. The criterion were (1) goodness of fit, (2) normality of deviations from the fit, and (3) equality of variance of deviations from fit for all crack lengths.

Probit, log-odds (linear) and arcsine models were based on $x = a$, without transformation of crack length scale. None of these models provided adequate goodness of fit in the sense that their patterns of deviation were not randomly distributed about the model over the entire crack-length range. Out of the other four, log-odds (log scale) model was consistent with the assumption of normality of deviations in most data set cases. There fore it was concluded that the log-odds (log scale) model provided an adequate fit to the POD(a) function for the "Have Cracks" data. This model was adopted for the simulation study.

Standard statistical regression methods can be used to fit the curve and also find a lower confidence bound on the true POD(a). The formulae for A and B are

$$B = \frac{\sum x_i y_i - \frac{1}{N}\sum x_i \sum y_i}{\sum x_i{}^2 - \frac{1}{N}\left(\sum x_i\right)^2}; \qquad A = \frac{\sum y_i}{N} - B\frac{\sum x_i}{N}$$

The formula for lower confidence bound on the mean for a given value is

$$m = A + Bx - t_{(n-2),g}\sqrt{\frac{1}{N-2}\sum\left(y_i - A - Bx_i\right)^2}\sqrt{\frac{1}{N} + \frac{\left(x - \frac{1}{N}\sum x_i\right)^2}{\sum x_i^2 - \frac{1}{N}\sum x_i^2}}$$

Where $t_{(n-2),g}$ is the γ^{th} percentile of the t distribution with (n-2) degrees of freedom.

The simulation investigation was based on 100 numeric repetitions of an experiment, each with a set of 400 inspections of details with different crack lengths. The data were analyzed by different methods including those based on binomial distribution and regression analysis using the log-odds model. Their conclusions were as follows. All cracks of the same length do not have the same detection probability. The POD as a function of crack length for the total population of structural details is a curve through the averages of the detection probabilities for all cracks of the same length (regression function). The log-odds model was acceptable as a regression function of the crack detection probabilities that were observed during the "Have Cracks" program. Confidence Limits (CL) can be placed on the POD function using regression analysis methods. The regression estimates of POD/CL are closer to the true POD, exhibit less scatter, and always provide an estimate of desired limit.

1988 Statistical Evaluation of NDE Reliability

Hovey and Berens reviewed the statistical tools in use by the industry during the late 80s [Hovey 88]. They describe the state-of-the-art analysis methods through examples from the retirement-for-cause (RFC) inspection system evaluation, with data collected on titanium web bore specimens. The log-odds analysis and the $\hat{a}$ versus a analysis had

become standard procedures by then. These two methods differ in the type of data they are designed to handle.

In many NDE systems, response signal amplitude is used for flaw detection decisions. The response signal is referred to as $\hat{a}$ to distinguish it from crack length *a*. If $\hat{a}$ is larger than the detection threshold, the system gives a detection indication, while $\hat{a}$ values smaller than threshold are ignored. A method of estimating POD function from $\hat{a}$ versus *a* assumes a linear relationship between $\ln(\hat{a})$ and $\ln(a)$ and normal distribution for the scatter in $\ln(\hat{a})$ about the mean trend. This is graphically presented in figure 9.1.

The response signal is assumed to have a lognormal distribution; and since the mean response signal is a linear function of log crack length, the POD function has the form of a cumulative lognormal distribution function. The logistic function is a very close approximation to the cumulative normal distribution function. The $\hat{a}$ versus *a* model therefore provides a theoretical justification for using the log-logistic distribution function to model the POD function. The log-odds function as seen earlier is analytically simpler and is thus a preferred choice for the analysis of find/miss data. Since the $\hat{a}$ data can be easily reduced to binary responses, the pass/fail analysis can also be used when response amplitudes are available.

Hovey and Berens performed the two methods on the RFC data and found that the mean trends for the POD function are similar for the two analyses; however, the confidence bound for the pass/fail analyses is much broader than the $\hat{a}$ versus *a* confidence bound. The broader confidence bound is due to the loss of information when reducing the response signal to one of the two outcomes - find or miss.

In their review, false calls are important from the operator evaluation viewpoint; and the two main tools are Coefficient of Contingency (CC) and the ROC curve. Both of these tools measure the tradeoff between POD and POFA.

Substantially more mathematical treatment on these two methods as applied to data sets is presented in [Berens 88].

1988 POD/SS

POD/SS was a set of PC-based programs for analyzing the results of NDE reliability experiments [Berens 88a]. It consisted of four programs – DATA (for data entry), PF (for pass/fail type of data analysis), AHAT (for $\hat{a}$ versus *a* type of data analysis), and FRTPRINT (for printing of output). Karta understands that UDRI has a much more recent version of POD/SS nearing completion that is based on a Microsoft Excel spreadsheet form and believed to be far more convenient and user friendly.

1989 Comparing POD Curves

Annis presents two statistical methods: statistical analysis of variance (ANOVA) and Chi-square to compare POD curves for quantification of system to system variations. He demonstrates the procedures using examples from ET data [Annis 89].

1992 Statistical POD Model using Actual Trial Inspection

The underwater NDE Center, UK, produced experimental POD curves for underwater inspection for fatigue cracks in tubular welded joints by comparing the results of underwater inspection with detailed in-air characterization of the cracks [Rudlin 92]. The curves produced were plotted as a function of crack length with depth information in the form of thresholds. Evaluation of two mathematical models (Logit Model and Exponential Model) of the experimental POD results showed that crack depth was the main contributory variable to the POD curves. Comparison of the models with data from trials on underwater magnetic particle inspection (MPI), two eddy current methods and ultrasonic creeping wave showed a reasonable agreement.

1996 Statistical Methods in NDE

Olin and Meeker provided a very good reference point for statisticians working in the area of NDE [Olin 96]. The document also carries peer review and discussions by Berens, Coleman and Ramsey, Spencer, Sweeting, and Tucker.

The statistical models are classified as those based on hit/miss data and continuous response data. POD curves can be plotted using the hit/miss data by either the logistic-regression model (or the log-odds) or the model free estimates (based on grouping of data into size categories). Generally these data are considerably less informative but convenient as compared to the continuous response data such as $\hat{a}$ vs. a.

The authors provide a good discussion on use of POD and ROC for evaluation and comparison of NDE flaw detection systems. POD describes inspection capability. Inspection intervals can be established based on prediction of fracture mechanics model to grow a 90/95 crack to failure. ROC curves are particularly useful in comparing different NDE methods and for assessing trade-off in choosing a threshold. The parameters to compare can be area under the curve or the distance from the point (0.5, 0.5) to the ROC curve.

The discussion section states that Coleman and Ramsey and Tucker recommend use of ROC, Berens and Spencer prefer POD with some provision for false alarm, and Sweeting recommends use of Probability of Indication [POI = POFA + (1-POFA) * POD].

1998 Fitting POD Curves to Hit/Miss Data

Analysis of hit/miss data should incorporate the concept that the data may have resulted from a process that produces hits or misses in a manner independent of flaw characteristics [Spencer 98]. This statement about the total process includes the human operator. The process can be done effectively by generalizing the usual POD curves to allow lower asymptotes other than 0 and upper asymptotes other than 1. The proposed form is

$$\mathrm{POD}(a) = a + (b - a) \cdot F(a : m, s)$$

where $a \geq 0$ is a lower asymptote, $b \leq 1$ is an upper asymptote, and $F(\cdot; m, s)$ function is one of the usual two parameter distribution functions (e.g., log-logistic or log-normal), that is a fit to the hit/miss data. The form indicated that fitting of a and b can result in an overstatement of the slope of the POD curve. A suggested approach to putting confidence bounds on the values, such as 0.90 detection crack length, would be to explore the distribution of that value as the value of a and b are systematically varied from their estimates back to 0 and 1 respectively.

9.10 Observations

- Statistical methods govern the design of the experiment.
- The POD curve represents NDI reliability in terms of detection only. This limitation is important from a safety viewpoint.
- The ROC curve represents the NDI reliability in terms of both detection and false calls. This information is important from a safety as well as an economic viewpoint. CC is another parameter that provides an equally significant measure of reliability.
- The best of the statistical procedures for data analysis today are the log-odds model and the $\hat{a}$ vs. a method. The former is good for hit/miss type of data acquisition, and the latter one is better when supported by a continuous function type of NDI response. Recent version of POD/SS should be a good tool to generate curves out of the acquired data for near future programs.
- Different statistical procedures are likely to provide differing POD curves. This variation poses a challenge in trying to present the results at the end of the assessment effort.

Best Procedure: Log-odds

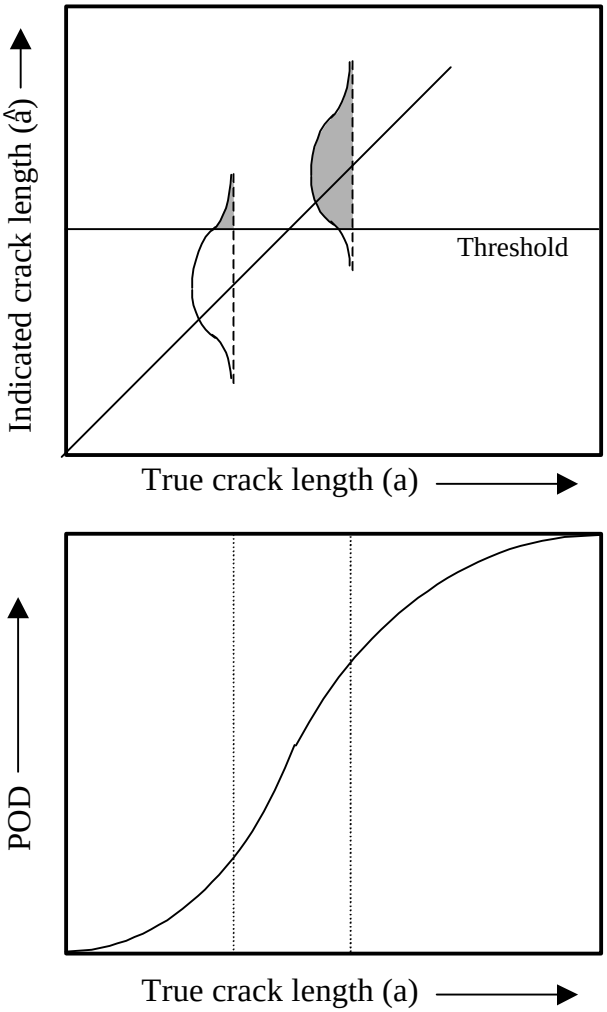

Figure 9.1: The $\hat{a}$ versus a to POD conversion. *A distribution of detection probabilities occur at each crack length. The scatter in this distribution is caused by the non-reproducibility of all factors other than crack length and is modeled as log-normal. The POD is essentially the probability that the indicated crack length exceeds the threshold of detection*

10. HUMAN FACTORS

When an NDI method fails to meet expectations, human factors are often cited as the sole cause. Experience has shown, however, that failure in implementation is most often due to failure to define or meet NDI engineering requirements. Though human skills development in pattern recognition and discrimination is an important element of most NDI procedures, human skills cannot compensate for shortfalls in NDI process control. Reliable NDI requires total attention to NDI process and to NDI processing criterion.

10.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

The probability of detection data had substantial variance among individual technicians that could not be explained by skills level, education, formal NDI training, or age. The variance was believed to be because of human factors, which was beyond the scope of the investigation. One of the human factors observed in the program but not addressed fully was the false call level. At that time (1978) no method was generally agreed upon for analysis of the impact of false calls on NDI reliability.

1979-84 NDI Technician Proficiency Program

The report makes no mention of any attempts to study human factors.

1987-88 Engineering Services in Support of NDI

This report covered procedures in support of NDI operations and did not address any human factors investigations of any sort.

10.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

The overall performance level of an NDI operation is dependent on the adequacy of the NDI engineering and acceptance criterion definition, the NDI materials, equipment processes, and human skills applied to the operations. The earlier study "Have Cracks" cited human factors as the primary factor for performance variation. Although human factors were observed, other inspection variables dominated the ALC facilities assessments. The ALCs observed that the most frequent cause of unreliable NDI performance was improper nondestructive engineering. In many cases, the NDI method selected was incorrect or was not qualified and controlled to the level necessary to obtain required discrimination. In other cases the NDI equipment, materials, and processes were

not controlled and did not provide discrimination required to the level qualified. In short, unless the NDI engineering is under control, the human operator at the end of the line does not have a chance.

Human factor variables observed were those due to knowledge, skills development, and experience in penetrant processing and in readout and interpretation of indications. Automation was often proposed to eliminate human factors. But without proper NDI engineering, specification of criteria, process validation, and process control, automation may result in improved process consistency but may not improve either the process capability or process reliability [Rummel 84].

10.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

This program gave no direct mention of any human factor aspect in the process assurance methodology developed. The field trials revealed distinct bimodal distribution of technician proficiency. Almost 60% of the airmen could not participate in a meaningful manner, as they were unable to independently perform the actions appropriate to the proper examination of various test sets. The difficulties were believed to be a direct result of insufficient training and independent experience in using basic NDI methods and equipment.

10.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

The SwRI did not directly address human factors. However, out of the 51 specific areas of concern brought out by the investigation, a good fraction of them related to human performance and man-process or man-machine interaction. Training needs and processes were greatly emphasized pointing towards human factors. Concerns were raised on the effect of work environment on human performance, NDI tasks not being designed with human performance in mind, negative side effects such as boredom, lack of feedback effecting motivation, lack of confidence, and ease of equipment usage.

1986-88 NDI Personnel Proficiency Evaluation Using UT

No human factors were addressed.

1990-94 Reliability Assessment Kit

This program developed a reliability assessment kit, and human factors study was out of the scope [Goodlin 94].

10.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

Extensive information on inspectors' physical and psychological factors was collected during the ECIRE. Some of these factors were later correlated to the observations of the experiment; however, no clear conclusions emerged from the project. Much of the differences observed between facilities and inspectors were attributed to differences in procedures.

10.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

Human and environmental factors were not included in the POD study. The exercise, however, did focus on procedural and process variables that required inspector input or adjustment during the course of study.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

Human and environmental factors were not included in the POD study.

10.7 NRC-Battelle

The NRC sponsored extensive work with a focus on human factors in NDE.

1985-86 Mini Round Robin Assessment of UT Performance

Battelle studied various personal, social, and organizational factors during the program. Technicians generally felt that the supervisor was of some use in determination of existence and size of the crack. Technicians also perceived that their supervisors were more concerned about a "false call" than a "miss." Organizational climate in terms of the way people work together and cleanliness at the work area effects performance. Technicians expressed that having to work around the schedules of others made it more difficult to do a good job.

The performance observations during morning and evening tasks were not significantly different. Fatigue caused by long work days primarily effected the far-side crack detection tasks. During the interviews, technicians grouped the fatigue in two categories: (1) fatigue due to long hours (usually exceeding 14 hours) combined with full protective gear, which leads to incident errors/mistakes and (2) fatigue due to working continuously over several weeks without a day off, which effects the attitude and general disposition. They expressed that fatigue often contributed to a wrong call or safety violation.

The characteristics that technicians considered to be very important in order to excel in performance were: (1) ability to concentrate, (2) understanding of NDT theory, (3) patience, (4) tolerance of environmental conditions, (5) manual dexterity, and (6) mathematical ability.

During interviews, several technicians indicated that neither the EPRI final examination score nor the certification level of a technician were strongly related to crack detection ability. Several of them indicated that the technique for passing the test was unrelated to the way that they would make decisions during an actual inspection. The difference was that during EPRI demonstration tests, emphasis was more on detection whereas in actual inspection, penalties for false calls are higher. Formal NDT certification was not seen as a good indicator of ability.

On the equipment side, various controls and displays can effect human performance. They (1) are often too small and too closely spaced for manipulation with gloved hands, (2) are not guarded against accidental actuation, (3) have small and difficult to read scales, (4) have screen displays that reflect glare, and (5) have a too small screen size. Technicians generally indicated a loyalty to a brand and model with which they are familiar.

Environmental conditions such as heat, humidity, noise, and lighting were of concern to the technicians. Protective clothing and gloves make it harder to move around. Technicians were aware of the radiation hazards and expressed that it did effect their work in some way. It leads to inspection in a hurry, shift of attention from task to personal safety, difficulty in seeing the equipment clearly, and difficulty in handling small search units with bulky protective gloves. These issues stem from the use of UT in a radiation environment.

Lack of feedback on how well technicians do, affects their performance, because it gives them an impression that utilities do not take their efforts seriously and does nothing to adjust their expectations.

1985-86 Human Reliability Impact

The emphasis in this review and analysis of human performance in NDT has been on difficult discrimination tasks, such as the detection and quantification of small intergranular stress corrosion cracks. The review showed that the ROC analysis method is the only approach available for separating the operator's discrimination ability from the operator's decision criterion. ROC analysis uses a statistical sampling approach that requires the use of a large number of material specimens, demanding substantial time and money.

The study suggested that the NDT technician performance could be improved through systematic application of human factor principles within the framework of ROC analysis. This was especially true in the areas of task and training variables and, to a lesser degree, in the area of procedural variables. Most of the variables and their impact were presented in the other complimentary report discussed above.

In another report [Spanner 86], which summarizes the program, a conclusion stated that no single performance-shaping factor (task, training, procedural, environmental, or individual) was responsible for the wide performance variations (unreliability) that had been observed. Hence improvement of any single factor will not have sufficient impact to correct this problem. However, if key human performance factors are all simultaneously adjusted in a positive direction, the resulting synergistic improvements in actual performance would be significant.

10.8 MIL-STD-1823 (Draft)

Not applicable.

10.9 Other Published Work

1974 Human Factors in NDE

Human Factors were identified as an important consideration early on by General Dynamics [Herr 74]. A very clear and simple article using examples pointed out issues and means to overcome effects of human factors. Selection and training of personnel must be the beginning of a reliable NDE program. Fracture critical parts should be identified to the inspector so that he can treat them differently. The procedures should be proofed by the inspectors under working conditions, and managers should not permit deviations. Higher wage classification for trained inspectors can help retain qualified personnel. Well-defined accept/reject criterion reduces the subjective decisions that an inspector must make. Environmental factors such as temperature, light and noise impact inspector ability. Inspections performed in awkward and cramped positions have a high fatigue factor with a corresponding lessening of confidence in inspection results. A companion subject to the environmental issue is management's attitude. Better results can follow if NDE is viewed by management as an integral, important part of the product. Reference standards for equipment and calibration procedures have a marked effect on an inspector's ability to detect defects and report data quantitatively. Automation for specific applications was suggested to eliminate human factors.

1985 USAF Reliability Programs

Petru [Petru 85] very nicely summarized the two major USAF efforts on reliability assessment of airframes and engines. Correct NDI engineering and effective process controls are indeed more significant than the human operator in achieving a reliable inspection. In other words, if engineering of the technique is poor, the operator has a little chance of success.

1985 Human Factors Considerations in NDI

Ainsworth reviews human factors issues in NDI and concludes that in many situations, it will be a human being in the system, rather than underlying technology, which will prove to be a limiting factor [Ainsworth 85]. The author was absolutely convinced that operator performance on most NDI tasks could be substantially improved by paying careful attention to the human factors involved, and then redesigning the tasks so that they are better suited to human abilities. Redesign of equipment, redesign of inspection procedures, changes in task design, better initial selection, and improved training methods can all play their part. In addition to the task designed with human factors consideration, they should be evaluated to determine whether under field conditions operators are capable of achieving the criteria demanded of them. Occasionally the new sophisticated techniques are less than effective because the operator has been virtually forgotten during the task design.

1987 Human Reliability in NDE

Glasch reviews the efforts on NDE reliability assessment prior to 1986 in the first part of his paper [Glasch 87]. In the second part, he reports on an experiment at the Aircraft Systems Department at the University of Illinois, Urbana Champaign. The purpose of the experiment was to compare the performance of novice inspectors with expert inspectors. Eight novice and three expert technicians participated in the experiment. All novices had formal training in NDE, hands on training in ET, and were students under 23 years old. All experts had more than 10 years experience in field NDE and were over 30 years of age. The specimens were a set of 16 aluminum brackets with four of them artificially cracked. Each subject was given 30 seconds to scan the 15x1.5cm area. The data were collected in the form of definitely cracked, probably cracked, possibly cracked, probably not cracked, and definitely not cracked. The task was performed in a well-lit classroom with no distractions from other people. For every confidence level, the hit rate was just as high or higher for novices as for the experts. Only by evaluating false alarm rates and plotting the information on a ROC curve could true reliability evaluation be performed. He reinforced the opinion of [Triggs 86] that POD and false alarms should be addressed together for NDE reliability assessments.

1989 Human Factors in Nuclear Power Plants

Within a conceptual framework, the worker is viewed as an open living system capable of self-direction and change. In this light the performance is not random. Rather, performance is purposeful and is organized by three aspects: (1) individual's goals, (2) behavioral and biological possibilities and limitations, and (3) context within which a person is functioning. A worker's final decision regarding a flaw may be shaped by multiplicity of internal and external factors. In particular, because of the central role that goal and context play in the organization of behavior, poor performance is not to be equated with personal deficiency [Behravesh 89].

A worker's poor performance can be best understood as being primarily a function of contextual and motivational factors. The contextual factors include the physical environment (e.g., machines, radiation, and heat) and the organizational environment (e.g., standards, attitudes, and practices of supervisors and managers). The motivational context includes a coherent purpose, high standards of professional conduct and the belief that effective performance is possible, and the availability of attentional and emotional energy resources. Those who are intrinsically motivated to do good work may be the most consistently effective workers.

Workers may perform poorly if they are unskilled in making decisions and solving problems in complex informational environment. The physical and mental stress associated with a job can quickly deplete available energy resources and lead to poor judgment. Anxiety and boredom can take their toll too. Finally in situations where informational and supportive feedback is not readily available, workers may develop negative beliefs about their capabilities and opportunities for exercising them.

Two separate studies were designed and conducted to identify relevant personal and environmental variables that influence performance. A series of semi-structured interviews of approximately one hour each were conducted with 37 subjects in the first study. In the second study, a sample of 100 episodes were generated from 20 subjects during the interviews. The main findings of the studies was that the supervisory and managerial practices were major determinants of the outcome of performance, assuming the workers have basic skills.

1995 FAA's Research on Human Factors in Aviation Maintenance

Shepherd and Dury [Shepherd 95] discuss the two broad methods for inspection reliability improvement. From the human factor side, the basic percept is that a mismatch between task demands and human capability leads to system-induced human errors. Explicit and quantitative models (e.g., decision theory, signal detection theory, and visual search theory) are used to predict the types of errors that can arise. This process then drives inspection system improvement; that is changing the task, the operator, and the machine or environment as appropriate. From the NDI reliability side, the basic percept is that all defects have a probability of less than 1.0 of being detected by real NDI systems, and that this probability must be predicted or measured for safe system operation. The models used are physical models of the structure, the defect, and the detection systems, with a degeneration factor due to human presence in the system. These models and the associated studies lead to POD and ROC curves. The process leads to reliability predictions, which drive the setting of safe inspection intervals. These two methods of enquiry may be presented separately, but they do have points of contact and overlap. In the outcome they have the same goals; that is, quantifying the performance of a man-machine system and predicting the effect of human and system changes on that performance. FAA's Visual Inspection Research Program combines the human factors and NDI studies.

1995 Empirical Approach to POD Modeling for MT

Lovejoy proposed an empirical model for estimation of POD based on controlled and uncontrolled factors [Lovejoy 95]. His model considered a fairly good number of parameters including definition of defects, technique, equipment, calibration, materials, performance checks, operator, work piece, location, and rate of inspection. Within the factor of human operator, he considered experience, training, eyesight, discipline, dedication, and concentration. To further properly account for the influence of various parameters, he intuitively associated weights with them. However, he made no attempt to justify the weights or provide any direction on how to obtain quantitative estimate of these parameters. As such, his intuitive mathematical model is of no practical significance.

1999 An Interview with an Ex-USAF Technician

Karta recently conducted an interview with an Ex-USAF technician who had performed NDI for the AF for almost four decades. Various human factors issues that effect the NDI performance were brought out. Excessive heat/cold effects the outcome of an inspection. Protective gear such as a flak jacket, web belt, canteen, rubber booties, helmet, and gloves restrict movement. Gas masks impair vision. It is very difficult to work in full NBC gear. Some inspectors have a tendency to continue using older equipment for reasons of familiarity and reluctance to learn new equipment. Age effects performance. Previous jobs and an intended career plan determine motivation to excel. If inspectors are pushed to complete the job ahead of schedule, false call and miss rate goes up. Emotions such as those caused by family, financial, and automobile problems play a big role. Performance at odd hours is never as good. Monotony of a task such as hundreds of fasteners or routine inspections that have never shown any defects makes an inspector go through the inspection with reduced attention.

10.10 Observations

- Major USAF programs did not give as much attention to human factors as they should have, in spite of the fact that fairly early they identified the need to do so.
- Much effort by various organizations and individuals has gone into identifying various factors that can influence an inspector's discrimination and decision-making abilities. These can really be classified into three categories: physical environment, organizational climate, and mental state.
- Physical environment involves such items as temperature, humidity, illumination, noise, posture, special protective gear and gloves. Most of these factors are somewhat controllable.
- Organizational climate involves management attitude, pressure of schedules, personnel behavior, hierarchical structure, and wages. These are also somewhat controllable.

- Mental state deals with intrinsic attitude and interest in the subject of NDI; internal motivation and zeal to excel; daily fluctuating factors such as mood, fatigue, sleeplessness, family, and financial situations. These are really uncontrollable.
- All these factors eventually impact an individual's decision-making ability at the critical moment of time.
- The most significant operator factor identified is that of the level and currency of skill development as applied to the inspection being performed.
- Routine and monotonous inspections with a history of predictable outcome make an inspector go through the inspection with reduced attention.
- A good number of investigations agree that if NDI engineering is not well defined, the human should not be blamed for poor performance.
- If inspectors are aware of a test, they can be expected to be more careful than normal production times.

Good Investigation : NRC-Battelle work

11. OUTCOME

This chapter summarizes results of various programs discussed in detail in chapters 2-10.

11.1 AF - Lockheed Georgia

1974-78 Have Cracks Will Travel

The overall reliability of NDI performed by the AF fell below what was previously assumed by established guidelines such as MIL-A-83444, "Airplane Damage Tolerance Requirements." The mean probabilities of detecting fatigue cracks in built-up aircraft structures, using typical maintenance inspection techniques and procedures, were at least 25% below the assumed values. The 90/95 reliability criteria could not be attained for any flaw size with typical inspection techniques applied by the average technician. With one exception, the NDI techniques employed in the program demonstrated considerable difficulty achieving a 50% POD with 95% confidence for 1/2-inch crack sizes. However, limited use of more advanced semi-automatic ET and UT (incorporated late in the program) indicated that 90/95 reliability criteria might be achievable at crack sizes somewhat smaller than the 1/2 inch measured by the program.

With one exception, the average capability among both field and depot NDI shops was found to be uniform. This aspect of uniformity was a strong point, which could be used to advantage if changes are incorporated AF-wide into the NDI system. Also, future measurements of AF NDI capability can be achieved by using a smaller sample of installation.

A distinctly higher level of flaw detection success was achieved at one exceptional installation, a depot, especially with the eddy current bolt-hole method. This demonstrated that considerably better performance levels than those generally exhibited are possible. This superior performance was attributed solely to individual proficiency.

The major variations in inspection results were found among individual technicians themselves. Surprisingly, formal education, technician age, skill level, classification, NDI experience, and training were found to have only minimal influence on performance levels. The primary source of variance among individual technicians was believed to be in the area of human factors, which was beyond the scope of this investigation.

One of the human factors observed in the program but not addressed fully was the false call level. At that time there was no generally agreed upon method for analysis of impact of false calls on NDI reliability.

The key recommendations of the program were: (1) concentrate on practical methods to evaluate the proficiency of the NDI technician as well as skill development and motivation of the technician; (2) evaluate technician proficiency through practical examination of flawed structures and not with commonly assumed indicators such as skill

level, education, years of experience, age, maturity, or hours of formal NDI training; (3) centralize NDI activities, and make NDI a full-time certified occupation, with a standard certification/re-certification program through administration of practical examinations at all bases and depots; (4) standardize the training program through development of training kits including actual fatigue-cracked structures and training manuals; (5) take advantage of evolving NDI technology; and (6) guidance from the user/technician, engineering specialists and personnel responsible for day-to-day operation should be included in programs whose objective is to develop improved versions of NDI equipment.

1978 Workshop

During August 2-4, 1978, eight task groups reviewed the data from the "Have Cracks" program and made specific recommendations [Lewis 78a]. They were:

1. *Personnel*: Individuals should be selected on the basis of identified traits, which are necessary for high proficiency and motivation.
2. *Training*: Classroom and hands-on training should be conducted at a centralized facility aimed at multi-skilled NDI capabilities.
3. *Certification*: All personnel should undergo a formal certification by examination.
4. *Management*: NDI should organizationally report directly to the chief of maintenance. Depots should monitor field operations and provide well-defined NDI procedures as Technical Orders.
5. *Equipment*: Improvements in equipment should be sought through automation and use of digital processors for control and data treatment.
6. *Reliability Measurement/Modeling*: Attention should be given to man-machine interfaces, physical/mental attributes, and equipment output stimulus level and patterns.
7. *Fracture Mechanics/NDI Interrelationship*: The trade-off between inspection interval and redundant NDI application should be examined for their impact on reliability.
8. *Data Analysis*: Further examination of existing data should be made before any additional data are acquired.

A recent understanding from the "Have Cracks" program essentially demonstrated that in-service 90% POD at a 95% level of confidence in a wide-area, “non-focused” inspection is not realistic. The program used a large population of flaws to get statistical validity, but suffered as a result. It is advisable to have narrowly defined critical sites identified in NDI procedures to assess 90/95 capability.

The depot that showed superior performance did not play by the ground-rules for assigning a random cross-section of personnel skills. The facility managers had a proficiency test in-force at that time and were aware of their inspectors' capabilities.

They offered their best personnel. Their selection however, showed the effectiveness of proficiency testing.

1979-84 NDI Technician Proficiency Program

Lockheed developed a test plan with specimens and a set of instructions that included a technician briefing, administrative procedures, NDI procedures, and other information related to data treatment. This plan resulted in technician proficiency evaluation and an opportunity to re-examine NDI reliability. Since the equipment standards and procedures used in this program were comparable to the "Have Cracks" program, the data indicated an improvement in mean POD that may be attributed to the differences in the structure inspected or improved technician proficiency. Knowledgeable observers agreed that whatever the effect of structure configurations on mean POD, technician performance did seem to improve. Improvement was attributed to the corrective measures taken in USAF training programs following the revelation made by the "Have Cracks" program. Continued effort to improve the POD and reliability for the NDI methods in use through better technician performance, better equipment, and better procedures seems the most effective way [Sproat 84].

1987-88 Engineering Services in Support of NDI

The program provided a detailed test plan for measurement of flaw detection capabilities of NDI on standards (specimens) and simulated aircraft structures in AF field and depot environments. The test plan contained procedures for program administration, NDI of the standards, verification of flaw response, and evaluation of the NDI process. Five NDI techniques - ET, MT, PT, UT, RT could be evaluated using NDI standards. The NDI standards consisted of three metallic standards (fastened aluminum joints and lugs and flat plates) and two composite standards (honeycomb assembly and transition joint).

Lockheed clearly described a set of specimens to evaluate eight different combinations of NDI procedures and specimen types including bonded attachments, composites, and honeycomb sandwich materials covering fatigue and delamination types of damages and the detailed NDI procedures.

11.2 AF - Martin Marietta

1979-84 Engine NDI Reliability

Martin Marietta identified several process control improvements and made recommendations to improve overall process control, inspection capability and reliability. Recommendations included some of made by direct observations in the production line and some as a result of off-line analyses. The team documented and described methodology, observations, and data processing so that performance improvements could be quantified at a future date. The documentation also was done so that the lessons learnt

could be useful in similar future studies[7]. During the program, performance results were directly reported to the responsible operational personnel for action.

Martin Marietta initially observed major variations in inspection performance in the production lines. They attributed variations to processing materials, process applications, equipment capabilities and equipment operations, and human factors. Wide variations in performance prompted the initiation of an extensive off-line program to relate samples of penetrant processing materials to production line performance.

Direct observations were that: (1) inspection materials were used without acceptance testing, (2) inspection equipment varied between facilities and within facilities, (3) automated equipment had little capability for adjustment to inspection process requirements, (4) inspection process and process applications varied, and (5) inspection operator training, instructions, supervision, and performance were variables.

11.3 AF - Battelle

1986-89 Surveillance and Control of AF NDI Labs and Shops

The most significant conclusion by Battelle from initial AF NDI field lab visits was that compliance checks and other measures were passive in nature. Discrimination of individual capabilities would require an approach in which the individuals were actively involved in the process (e.g., inspecting a test sample). The other significant observation was that there are logistics issues associated with doing NDI process assurance as approximately 100 active labs and 900 active NDI practitioners are scattered throughout the continental USA, Europe, and Far east.

During the second part of the study, Battelle evaluated two concepts: (1) having a process assurance representative from a central authority, (e.g., NDI Program Office) or the command level NDI manager), go to the field labs and (2) having an individual NDI practitioner from field labs come to a central site, (e.g., NDI Program Office). Battelle evaluated implementation scenarios in terms of the time and cost associated with processing a given number of airmen/labs in a given period of time, typically a year. The most promising course was that in which one person from a central authority visited the field labs and performed the process assurance on site. These would be one-day visits in which five of the lab's staff would be measured in each of the five NDI methods. With a reasonable travel schedule, it was possible with this approach to measure approximately 90% of the labs and 50% of the active-duty AF NDI practitioners in a year at an annualized cost of $112 per airman.

In the third part of the project, a preliminary 2-day trial was held at Pease AFB. The results were encouraging but indicated a need for a simple T.O. like instructions for each

[7] The fact that today we are reviewing these efforts as a pre cursor to protocol program development fulfills the original secondary objective conceived by AF in early 1980s.

test set and easier to examine UT and ET test sets. (The original ET and UT test sets were from technician proficiency kits) These changes were made, and more extensive one-day field trials were made at four labs with sixteen practitioners. The results included following items: The field trial revealed a bimodal distribution of technician proficiency, (i.e., they formed two clusters). Almost 40% of the technicians fully participated and demonstrated proficiency in one or more of the NDI methods. The remainder could not participate in a meaningful manner as they were unable to perform the actions appropriate to proper examination of the various test sets. Particular difficulties with this group included setting up and calibrating the UT equipment, manually handling the penetrant and MT samples so as not to destroy the indications, and properly aligning the X-ray test piece with the incident beam. These difficulties were believed to be a direct result of insufficient training (especially on the job training), and insufficient independent experience using basic NDI methods/equipment. Another result of the field trial was the demonstration that a moderately proficient inspector could complete on average one method per hour. This meant that five practitioners could be measured in all five basic NDI methods by a one-person, one-day visit to a field lab.

The project concluded with the preparation of a detail process assurance methodology based on one person, one day, five methods, and a five-airman demonstration of individual proficiency by means of the use of well-characterized test samples. Statistical procedures developed on other NDI projects were adapted and included with the preferred approach.

Although not an objective of the project, one consequence of the field trial was the revelation that the general level of demonstrable NDI proficiency in the field labs was much less than anticipated.

11.4 AF - SwRI

1987-88 Recommendations for AF-NDI Technician Proficiency Improvement

During the course of the project, SwRI initially identified 277 areas of concern including duplicates from different sources. These were then reduced to 51 specific concerns and three issues, under eleven general categories. These 54 concerns were then rated based on criticality. From these, SwRI generated 227 candidate solutions to address these 54 concerns. These solutions were combined and condensed to form one or more recommendations for each area of concern for a total of 75 recommendations. These recommendations were then presented to a panel of experts to achieve a group estimate of the potential promise, feasibility and cost. The process used multiple-pass evaluations using a modified Delphi approach with a subject matter expert from outside the program acting as a moderator. Each expert rated all the recommendations without discussions among each other. The disputed ratings were reassessed by open discussion. Finally, the recommendations were rank ordered within each of these general categories, according to

the sum of their ratings for promise and feasibility. SwRI also computed and reported aggregate ratings for promise, feasibility, and cost.

SwRI made specific recommendations addressing the areas of concern. The global areas of concern were in terms of measurement of NDI proficiency, role of NDI personnel, and responsibility of NDI personnel. The report states "The fact is that we were currently unable to assess NDI proficiency." The specific areas of concern were the selection of NDI personnel; initial, on-job, and continued training of personnel; work environment; performance; procedures; and equipment. Several of the recommendations could be combined to address more than one area of concern. For example, a computerized trainer/evaluator could address a large number of concerns such as initial training, on-job training, performance measurement, feedback, standardization, and motivation. The solution may appear expensive, but it is most cost effective.

1986-88 NDI Personnel Proficiency Evaluation Using UT

In Phase I subset and Phase II, almost half of the technicians qualified as proficient technicians. The procedure for technician proficiency evaluation was demonstrated to be feasible. Technician performance appeared to be less dependent on detection criterion than on individual detection capability. Technician proficiency can be evaluated accurately with a high degree of confidence using CC and ROC curves. False percentage data (POFA) appear to be the deciding factor in eliminating a number of individuals from the acceptable proficiency limits, regardless of high POD.

1990-94 Reliability Assessment Kit

SwRI developed, validated, and provided the USAF and Naval Aviation Depot with a kit for determining the reliability of specific NDI methods used for evaluating airframe structures. Testing procedures and standards were developed for RT, ET, MT, and PT, UT. The test kit was validated at SA-ALC. Inspection setup and care procedures controlled the examination of the standards in a manner similar to the T.O.s. The standards included in the test kits were designed to simulate characteristic airframe structures. They contained a wide range of deliberately induced characteristic defects of varying sizes for the purpose of determining the POD for such defects. Each test kit included the statistical data entry and analysis software.

11.5 FAA - Sandia National Labs and SAIC

1992-94 Eddy Current Inspection Reliability Experiment (ECIRE)

The data gathered during the ECIR experiment indicated that the field factors, in aggregate, effect performance levels. From the laboratory inspections conducted, the 90^{th} percentile point on the POD curve for crack inspections on bare surface was estimated to be within 60-70 mil range. The same percentile as an average from the field data was

approximately 90 mil. Inspector-to-inspector differences were a major source of variation in inspection results. Facility differences were significant. Other factors affecting inspections included surface conditions, crack orientation, and accessibility. Procedural tedium was not a significant factor. Individual variations in performance were not explained by uncontrolled factors such as age and recency of experience. Much of the variation observed was attributed to differences in procedure. Departures from the Boeing developed procedures did not perform as well. Several facilities used two-man teams but no significant improvement was observed. Although the inspectors were trained in ET, many did not have equipment-specific training. This lead to less-than-optimal settings and difficulty in making calls. The program successfully characterized the ability of industry inspectors to detect cracks using their own equipment and procedures.

The recommendations of the experiment were: standardize training by providing minimum training standards and training certification requirements, encourage the use of more modern, sensitive instrumentation, along with adequate training in its use; and ensure that the procedures for calibration, set up, and use of eddy current instruments are clear, understandable, and usable.

11.6 AF - SAIC

1996-98 C-141 Lower Wing 2nd Layer Inspection

Both the lab and field inspection data obtained during the SAIC program indicated that the UT procedure for C-141 wing splice for second layer cracks was capable of finding the target crack of 0.05 inch. The overall 90% detection crack length was 0.073 inch with a false call rate of 9%. Expected variations in the setup parameters were shown in the laboratory to not only influence the characteristics of the signal, but to do so enough to cause variations in the calls made from those signals. Analysis on the common signal data set indicated that much of the inspector variation was due to different signal interpretation. All of the objectives initially defined were met.

1998-99 C-141 Lower Wing Simultaneous 1st and 2nd Layer Inspection

The field results demonstrated that the first-layer detection capability could be added to the procedure already developed for second-layer fatigue-crack detection. The data supported the earlier derived estimate of a 90% detection rate for 2nd layer cracks with a length of 0.073 inch for identifying a rivet with a flaw. However the rate was only accomplished when credit was given for a find as long as a call was made even if the layer in which it was located was misidentified. Notches of 0.053 inch in the 1st layer were estimated to have a POD of 90%. The value was based on an inspector correctly identifying the layer as well. If credit were given for any identification of a flaw irrespective of the layer, then a 90% detection flaw size was estimated to be 0.040 inch.

1998-99 C-130 Center Wing Stringer - Hat Section

Estimated capabilities for C-141 second layer inspection held good for C-130 second layer inspection as well. No major adverse impact occurred from the added geometry conditions.

11.7 NRC-Battelle

1985-86 Mini Round Robin Assessment of UT Performance

Level II technicians performed no better or no worse than level III technicians; however, no specific conclusion was drawn because the number of level II technicians was only four. Fatigue had no statistically significant effect on performance. Detection performance for near-side long cracks was no better than near-side short cracks. Far-side detection was essentially no better than chance. None of the UT equipment was judged to be totally acceptable on the basis of its human engineering design characteristics. ROC analysis works exceedingly well for analyzing technician performance.

11.8 MIL-STD-1823 (Draft)

The outcome of this program was a document on recommendations for conducting NDI reliability assessment experiments with very valuable information. After staying in the draft form for 10 years, it was accepted as MIL-HDBK-1823 in April 1999.

11.9 Other Published Work

Nothing relevant was found.

11.10 Observations

- All the programs appear to have met their objectives.
- The early programs revealed that field NDI is not as reliable as originally perceived.
- Subsequent efforts to perform technician proficiency evaluation and improvements did help make NDI more reliable, but the need for such programs continues to exist.
- Human factor investigations did not receive as much attention as it deserved.
- Training has been a fairly common recommendation from most of the programs. Skill development and skill revalidation provide significant opportunity for improvements.
- Any attempt to assess the best achievable NDI performance could not be seen.

- Battelle work created a detailed process assurance methodology and tested it in field (1989). No subsequent use of their proposed process has been traced.
- SwRI developed and validated a reliability assessment kit (1994) including the specimen hardware and data analysis software. No report is on record on how and when the kit was used[8].
- Substantial effort and experience have gone into MIL-STD, but the document's direct usage appears to have been very limited.

A Very Good Program: FAA

[8] Though, Karta knows from one of the program participants that it was used for ET POD data generation.

12. OTHER NDI RELIABILITY PROGRAMS

So far the reports and investigations from various major reliability assessment programs and some of the related research activities have been reviewed. Many other minor NDI reliability assessment efforts are also in the literature. Some of these reports that were obtainable are reviewed in this chapter.

12.1 US Air Force

1976 AF - Vanderbilt University

Packman et al. appear to have been pioneers of the NDI reliability assessment program. The Air Force Office of Scientific Research supported their work assisted by NASA and General Dynamics. Most of the original definitions and explanations can be seen from their reports [Malpani 76, Packman 76]. Their reports define the meaning of 95% POD at 90% confidence as: "If 100 sets of parts are inspected, and each set contains 100 flawed parts (a grand total of 10,000 flawed parts), there is a probability that at least 95 of each 100 flawed parts will be detected in at least 90 of the 100 sets of inspection. No information is conveyed regarding the actual number of flaws the 100 sets of inspections will find."

Considering that their work is almost at the beginning of the NDI reliability assessment era, the demonstration program as proposed [Malpani 76] was very mature and well thought out. The core of the experimental procedure as described by them has not changed since then. It has only been refined over the last 25 years in terms of direct details. The significance of accounting for false call was quite well identified. Variability in reliability due to inspector has been mentioned. The statistical tools, though limited, were probably the state of the art at that time. There was no way to circumvent the basic statistical requirements if a specific combination of reliability and confidence level were demanded. They present the sample size requirements for specific combinations as

POD/Confidence	Success/trials								
90/95	29/29	45/46	59/61	72/75	85/89	98/103			
90/50	7/7	16/17	25/27	34/37	43/47	52/57	61/67	70/77	79/87

They recommended mixing of flawed and unflawed specimens in a ratio of 1:1. The details of the data analysis methods are not discussed here, as most of it has become outdated now.

They evaluated data from three production NDI procedures: PT of Ti-6Al-4V plates using high-resolution penetrant process, PT of Al 7075-T6551 plates using medium resolution penetrant, and MT of high-strength D6AC steel plates. The 4x12x0.50 inch

plates were fatigued in a three-point bend fixture to produce cracks in a 0.008 - 0.025 inch range either from an initial laser or weld solidification spot, and reduced to 4x8x0.05 inches with the flaws randomly located on the surface. The specimens (50% flawed) were sent to the production inspection lines in groups of approximately 25-30. The specimens were cleaned and reused until a sufficient number of inspections had been made. The results based on their methods of estimation were: Ti PT 95/95 POD at 0.055 inch, and Steel MT 95/95 POD at 0.080 inch. The limitations of the data that were brought out were primarily related to difference in reality and experimental situations. The inspection procedures were extremely specific and carefully detailed, flaw types were specific, specimens were similar but not identical, and they were a simple shape and low cost. General Dynamics conducted additional studies and developed a model to translate the NDI capabilities assessed from these simple geometry specimens to equivalent detection sensitivity for specimens with complex geometries [Chang 76]. They developed an adaptive learning technique and a linear regression analysis to establish the parametric relationships between inspection sensitivity and NDI parameters.

1986 POD Estimation of Sub-surface UT

Acquisition of empirical POD data involved use of controlled samples of materials containing discontinuities of known characteristics. For subsurface inspection, developing such a database presented difficulties in obtaining or producing samples covering the full range of possible characteristics and in adequately identifying those characteristics with a "referee" technique. Sturges et al. described an equivalent reflectivity method whereby estimates for the POD of subsurface discontinuities are derived from an algorithm linking metallographic and UT measurements on a relatively small number of artificial and natural reflectors [Sturges 86]. This method was subsequently developed to add confidence levels to the analysis. During a discussion on the paper, Mr. Jan Van Den Andel suggested the use of models in glass for volumetric visibility and relatively easier defect characterization. Subsequent to the development of the $\hat{a}$ vs. a method, this procedure was modified to account for truncated and censored data; but required more statistical and data manipulation expertise [Burkel 96].

1995-98 POD of Hidden Corrosion

Engineers have nothing analogous to crack length to describe corrosion damage. For purposes of POD, percentage material loss (PML) is a possible index [Matzkanin 98]. In the Air Force Disassembly and Hidden Corrosion Detection Program, ARINC evaluated inspection reliability for detecting hidden corrosion using several different NDT techniques [Howard 95]. To determine corrosion detection POD, they fabricated lap joint specimens from aluminum sheet. To simulate the material loss associated with corrosion damage, they used electrical discharge machining to create square areas of thinning on the backside of the top sheet of each lap joint before assembly. They also designed lap joint coupons with specific flaw locations and flaw geometries to test the effects of flaw area, flaw density, and sheet thickness on POD. The POD functions were plotted by fitting the log-odds function to the binary hit/miss data [Berens 88]. They showed that

two-frequency eddy current procedures have good detection capability particularly above 10 PML. Large-area flaws were generally easier to detect, low flaw density contributed to better POD, and thinner sheets contributed to better POD. For eddy current procedures, flaw volume proved to be a better metric than flaw area or PML, particularly for smaller volume flaws where the flaw was much smaller than the eddy current field.

USAF tasked UDRI to optimize existing corrosion detection technologies and couple them with an automated system to provide significant maintainability improvements to the C/KC-135 and similar type aircraft [Hoppe 98]. Special emphasis was placed on the NDI evaluation and validation methodologies developed and applied in this program. The specimens acquired came form two different aircraft, a scrapped KC-135 previously stationed in Hawaii, and a B-707 fuselage a commercial version of KC-135. UDRI also engineered several other panels, each having material removed from one side and assembled into skin structures with simulated corrosion by products in the void. The specimens were designed to correlate the NDI output and the thickness loss. Conceptually, UDRI's method was based on the premise that for each NDI technology being evaluated, the output at a given point is a function of the thickness loss in a small region around the point (cell). An interesting observation on the POD curve was that it became absolute 1.0 at certain average material loss, unlike POD of cracks which hardly ever reaches 1.0.

1998-99 Reliability of MOI for C-5 Fuselage

SwRI conducted a program for the SA-ALC at Kelly Air Force Base (AFB) to quantify the reliability of the magneto-optic eddy current imaging (MOI) inspection system developed by PRI Instrumentation for detecting defects in the skin of the C-5 aircraft fuselage [Fischer 98]. The primary emphasis was determination of POD of first-layer cracks extending radially past the edge of the fastener head. Secondary objectives were to make limited assessment of the capabilities of the MOI technique for detecting corrosion and second-layer cracks. In order to provide realistic POD estimates, SwRI conducted experiments to identify significant variables, the range of test conditions, and conditions found in aircraft that adversely affect the performance of the MOI system. Variables found to be significant included flaw orientation (parallel or perpendicular to the first layer edge), fastener-to-edge spacing, fastener-to-fastener spacing, fastener head height, paint thickness, skin curvature, and MOI excitation frequency and power level. Variables found to be not significant for first-layer flaws were fastener diameter, first- and second-layer thickness, and second-layer geometry. A statistically designed experimental program was developed to determine POD in aircraft conditions as determined from measurements taken on an aircraft. SwRI fabricated specimens containing fatigue cracks according to this experimental design. They wrote procedures and trained Kelly AFB inspectors in the use of the MOI equipment and procedures. Four of the trained inspectors were given blind tests using the specimens. Separate tests were conducted with aluminum, titanium, and steel fasteners. The results were tabulated separately for each of the three different fastener types, and also separately for cracks that are connecting (that link two fasteners or a fastener and an edge) and non-connecting.

The Air Force criterion for acceptable performance was a POD of 90 percent with a 95-percent lower confidence bound for first-layer fatigue cracks 0.17 inch (4.3 mm) or more in radial length from the edge of the fastener head. The results were that this criterion could be met with aluminum and titanium fasteners for cracks that are non-connecting. For connecting cracks, the criteria were almost met for aluminum and titanium fasteners. For steel fasteners, the criterion was not met in either case.

As a result of this work, PRI Instrumentation made improvements to the MOI system to address the difficulties found, and the POD tests were repeated using inspectors from PRI and AF [Burkhardt 99]. Improved POD results were obtained. The difference between PRI and AF inspectors was found to be minimal in spite of the fact that PRI inspectors are better trained on MOI systems. The AF goal of detecting 0.17 inch flaw with 90/95 reliability was achieved for titanium fasteners and not for steel fasteners.

1998 POD Assessment using Real Aircraft Engine Components

Fahr and Forsyth reported experimental procedures and results of POD measurements on service-expired components from a military aircraft engine [Fahr 98]. They compared the service-induced low-cycle fatigue cracks present in the components with artificial cracks in terms of physical characteristics and NDI response. The experiment concluded that the results of manual ET were not as good as those from automated ET but better than PT, MT and UT. NDI response from artificially induced cracks could be quite different from the service-induced cracks due to shape, surface texture, and tightness. They recommended that for realistic POD measurements, actual parts with real service-induced cracks should be used if possible.

12.2 US Navy

1988-90 UT Vs. RT for Weld Inspection

A major advantage of RT compared to other inspection techniques is the availability of objective quality evidence as a permanent record of inspection. Reliability of manual ultrasonic (MUT), computer-assisted ultrasonic (CAUT), and RT was compared at David Taylor Research Center for the US Navy [DeNale 89, DeNale 90]. The specimen set included 16 welds with purposely induced discontinuities, and 17 welds removed from service. The welds were fabricated from steel in the form of double-V-groove joints, nominally 1.5 inches thick. The 33 welds contained 191 discontinuities of the following type: slag, lack of fusion, incomplete penetration, cracks, slugs, clustered porosity, and scattered porosity. Eight inspectors performed UT on 33 test welds, and eight interpreters reviewed radiographs of the 33 test welds. Discontinuities were identified by reviewing all of the inspection results, characterized as to type and verified by sectioning and metallography. The Navy used the information to document the following: the ability to detect discontinuities of specific types and sizes, probability of accepting or rejecting specific discontinuity types and sizes using current UT and RT acceptance

criteria, the repeatability of inspection methods, and differences in the rejection rate of a weld using MUT, CAUT, and RT. The conclusions were: (1) UT has a higher probability than RT of detecting and consistently rejecting planar discontinuities, (2) UT and RT have comparable capacities for detecting and rejecting volumetric discontinuities, and (3) UT is an acceptable alternative to RT for weld inspections from a reliability standpoint.

12.3 FAA

1995 Visual Inspection Research Program

Visual inspection of aircraft structures was a topic of FAA advisory circular AC-43-xx in 1994 and is a topic of great interest to aircraft manufacturers and airlines. FAA instituted the Visual Inspection Research Program to measure and improve visual inspection. For this program both NDI engineering and human factors were addressed as visual inspection is used to detect more than just well-defined cracks. Visual inspection is not even limited to vision, as active tactile (haptic) search supplements vision to detect corrosion and loose and worn parts [Shepherd 95].

1999 AAWG Action Item - Lockheed Martin

Since 1998 Lockheed Martin has been investigating the lower limit of inspection capability to detect widespread fatigue damage in riveted and fastener joints. Preliminary work on baseline inspection capability of two types of NDI equipment (UT and ET) to detect EDM notches in typical aircraft structure has been completed [Otterloo 99]. Lockheed used four types of specimens as defined by the airworthiness assurance working group. These included three-layer thin-skin structure with flush head rivets, protruding head rivets, shear head pin, and collar fastener. Highly experienced inspectors performed the examination. The information is now being used to fabricate fatigue cracked specimens. These specimens will be used for POD studies and further comparison of the notch and crack responses for eddy current inspection. The notch size detectability is expected to be smaller than the 90/95 fatigue crack detectability. The final part of the program includes a POD study on real aircraft structure under inservice inspection conditions.

12.4 Power Industry

1980 Beginning of the NDE Reliability Assessment in the Power Sector

Discussions among Mordfin and a few other representatives of the power industry offer the beginning of NDE reliability assessment in Power Sector [Mordfin 80]. Mordfin presented his views and need for establishing reliability of NDE inspection systems as

more critical than small but uncharacterized improvements in reliability. He emphasized development of a methodology for demonstration and verification programs that are based on standards, well-defined systems, rigorous procedures, and appropriate statistical analyses. He also prioritized research to better characterize the contribution of the NDE inspector to system reliability. The discussion panel agreed that reliability of NDE is a critical issue; however, a difference of opinion arose about what needed to be done. While Mordfin proposed standardized procedures, others suggested approaches towards increasing reliability through training, commitment, and automation. Since then the scientific community has come a long way, and clearly Mordfin had vision.

12.5 Observations

- All industries need NDI reliability assessment programs.
- Limited work has been done on POD data generation for corroded parts and sub-surface cracks.
- FAA has done work on visual inspection reliability.

Program of Significant Value: VIRP of FAA

13. DESIGN OF NDI RELIABILITY EXPERIMENTS

While reviewing major efforts and their results, Karta came across reports that have led to good programs and have the potential to provide direction to future reliability assessment programs. This chapter reviews manuscripts on design of NDI reliability experiments.

13.1 From Air Force-Sponsored Programs

1982 Guidelines for NDI Reliability on Aircraft Production Parts

Guidelines were prepared for ASNT and approved by ASNT National Board of Directors in 1976 on recommended practice for demonstration of NDI reliability on aircraft production parts. A form of this document was published in 1982, which contains information necessary for development of a valid, repeatable NDI demonstration program [Rummel 82]. It describes in detail the various operational requirements - general, equipment, personnel, specimens, inspection and reporting procedures, data acquisition and reduction procedures, and qualification and re-qualification procedures. Sufficient details are provided on the subject of operating parameters, statistics, and metallic specimen preparation. The document did not address the issue of flaw-size resolution once the flaw was detected.

This document was intended to be used as a guide for preparation of specific reliability demonstration plans. It was also projected as a living document that should continue to be a focal point for technology development in the aerospace community. Most of the subject matter is still useful; however, for a few good reasons, the document is not adequate in isolation. Advances in computational technology can reduce the cost and effort of reliability assessment programs. Advanced materials with their own specific damage types require different approach to specimen design. The need to identify false calls probably demands use of ROC over POD data.

Although this document is now considered obsolete [Easter 98, page 63] the methodologies for flaw generation, NDI application, and data recording remain as useful guidelines.

1988 Concerns in Design of NDE Reliability Experiments

Berens summarized very well four major concerns while designing an NDE reliability experiment [Berens 88]. These are: (1) the method of controlling the factors to be evaluated in the experiment, (2) the method of accounting for the uncontrolled factors in the experiment, (3) the number of flawed and unflawed inspection sites, and (4) the sizes of the flaws in the specimen.

An experiment designed to demonstrate the NDI capability assumes that the protocol for conducting the experiments is well defined for the application, that the inspection process is under control (hit/miss decisions are stable over time), and that all other factors introducing variability into the inspection decision will be representative of the application. The most important of the factors introducing variations are (1) differences in physical properties of cracks of nominally identical sizes, (2) basic repeatability of the magnitude of NDE signal response when a specific crack is independently inspected by a single inspector using the same equipment, (3) summation of all human factors associated with the particular inspectors in the population of interest, and (4) differences introduced by the changes in the inspection hardware. Berens makes a particular note that *k* inspections of *n* flaws is not equivalent to inspections on *n.k* different flaws, even if the inspections are totally independent.

1989 Design of Capability and Reliability Experiments

Capability assessment involves observing the condition of facilities, materials and equipment, operating practices, and overall expertise in execution of the NDI process. Reliability determinations measure flaw detection probabilities derived by NDI and the ability to discern flaw characteristics [Hovey 89]. In this manuscript, the authors discussed various details in connection with a proposed program for AF NDI capability and reliability assessment. They pointed out two main constraints of any such program. First, is the human factor. It is impossible to conduct an evaluation that the operators are not aware of, and their knowledge of evaluation will bias the performance of inspections. The evaluation program can at best simulate NDI in real situations as closely as possible. Second, is the economics. The POD function is different for each combination of material, geometry, flaw characteristics, and inspection method. The total number of combinations is far too great to evaluate in one program. Compromise is thus required.

The uncontrolled variables that must be accounted for in the sampling plan include human factors, inspection equipment, and flaw-to-flaw variability. Human factors and equipment variability are linked because the inspectors will perform inspections with a specific system that is familiar to them, thereby simulating day-to-day operations as closely as possible. The controlled variables of the sampling plan include flaw sizes and inspection procedures. The proposed sample size was 60 flawed specimens in the estimated 1 to 99 percentile POD range and at least twice as many unflawed specimens. Experiments need to have very precise instructions on their conduct and data acquisition. The procedures are not difficult to follow, however strict compliance is essential for a valid and representative statistical evaluation.

1989 Application of NDI Reliability to Systems

The success of a reliable NDI application depends greatly on the expertise and thoroughness of the NDI engineering that is performed. Most failures in NDI system applications and in the automation of an NDI system can be attributed to failures in NDI engineering and to unrealistic NDI performance expectations [Rummel 89]. This

manuscript, which is more like a refresher course in application of NDI, very comprehensively presents aspects of NDI.

1989 MIL-STD-1823 (Draft)

The experimental design defines the conditions related to NDI process parameters under which the demonstration inspections are to be performed. In particular, an experimental design comprises: (1) identification of process variables that may influence flaw detectability, but cannot be precisely controlled in the real inspection environment, (2) specification of a matrix of inspection conditions that fairly represent the real inspection environment by accounting for influencing variables to permit valid analyses, and (3) the order for performing the individual inspections of the test matrix. Although general guidelines for these areas are presented in the MIL-STD, it is recommended that a qualified statistician should participate in preparation of the experimental design.

The design of the experiment provides the foundation for the entire system evaluation. No amount of clever analysis can overcome a poorly designed experiment.

13.2 From FAA-Sponsored Programs

1993 Generic Protocol for Inspection Reliability Experiments (FAA-Sandia)

Reliability experiments require a functional plan and detailed plan [Spencer 93]. A functional plan spells out the goals of the experiment, methodology for creating the hardware, implementing the research program, observing human factors, data collection and analysis. The detailed plan lists the steps that must be taken to fulfil the mission and goals of the functional plan. These plans need to be backed up with a contingency plan for unexpected situation(s). A statistician needs to be involved during the detailed planning phase. This document [Spencer 93] is a valuable resource for NDI reliability experiment design.

The major goal of the experiment is usually to assess reliability of NDI systems under representative conditions of application. Specific secondary goals of a program may also include the quantification of the effects that certain variables have on overall reliability. The design must then address how these variables are to be incorporated into the inspections and how they are to be set during the experiment.

Experimental design begins with identification of test variables such as facility environment, human factors, equipment, and inspection process and procedures. Controlled variables have specified values for the experiments, and uncontrolled variables receive values of their own during the experiment. During the experiment the actual values of both the controlled and uncontrolled variables should be recorded. In an experiment where several variables are to be controlled, care must be taken to assign the values of each variables with each inspection in such a way as to assure that the factor effects can be estimated.

POD curves in the form of a linear log-odds model were shown to adequately fit the NDI reliability data [Berens 88]. In this model the natural logarithm of {POD/(1-POD)} is expressed as a linear function of natural logarithm of crack size(a). As is true for any regression problem, the estimated curve fit is more precise in the region of flaw sizes where data exist than it would be for flaw sizes outside the region of existing data. For this reason it is desirable to have the flaw sizes distributed over the region where the POD curve has a higher slope. A given set of specimens may be used in various NDI experiments with different equipment. The region of flaw size providing the best information for each experiment is likely to differ. To accommodate multiple uses of specimens and to minimize the chance of completely missing the range of interest for any one experiment, it is suggested that the flaw sizes be uniformly distributed between the minimum and maximum of potential interest. As log(a) is often used in the modeling of the POD, it is also reasonable to distribute the flaw sizes uniformly on the log scale [Berens 88]. Doing so would, however, result in a distribution that favors the smaller over the large sizes. Experience has shown that 30 flaws with flaw sizes covering the region of interest (10^{th} percentile to 90^{th} percentile) are usually sufficient. Because the correct region is not known in advance and because more precision is gained by more flaws, a minimum of 60 flaws should be considered in an extended range of sizes [Annis 89, Hovey 89, and Berens 88]. It is also of interest to determine the propensity of an NDI system (including an inspector) to make false calls. Protocol also recommends that the experiment contain about 3 times as many unflawed sites as flawed sites [Hovey 89, Berens 88]. In the discussion and reporting of the data, the experimental flaw density and its relationship to that density experienced in actual inspections should be addressed. These guidelines include each type of flaw in the assessment program if flaw characteristics other than size are believed to have substantial impact.

For estimating facility-to-facility differences, each facility represents a single data point regardless of the number of inspections taking place within a facility. The number of facilities to be included in the experiment should be determined in order to achieve a given probability of obtaining at least one extreme facility. Based on random sampling, it is as shown in the table below.

Total number of facilities		10	20	30	40	50
Sample size	90% confidence	9	14	16	17	18
	95% confidence	10	15	18	20	21

If information exists on likely causes of variation, it may be possible to choose "judgment" samples to reflect the variation and thereby reduce the number of required facilities. The number of inspections required at each facility should be such that no one inspector has to inspect more than once. If an inspector has to perform more than once, the time interval between two inspections should be large enough to minimize the chances of memory associated with the first inspection.

The specimens must not only model the structural points of interest, but also represent the global geometry as it normally presents itself to the inspector. However, a tradeoff is needed between cost/time and prioritized goals of the experiment. Building an entire

fuselage or wing section is a possibility, however, shipping and assembly logistics warrant that specimen dimensions be minimized. Cost of manufacture may inhibit producing natural specimens. A statistically desirable experiment with natural flaw distribution is likely to demand a larger number of specimens. Control on size and location of artificial flaws is easier to achieve than those produced by fatigue cycling. Experimental setup should be close to normal practice. Use of an actual airplane could be ideal, but it is very difficult to obtain the desired flaw density; and much of the experimental control is lost.

Specimen identification is key to information tracking of flaw characterization and performance evaluation. Identification should be transparent to the inspectors. If possible, capability should exist to ask for a re-inspection by a given inspector without that inspector knowing about it. Specimen characterization by independent determination of flaw dimensions should be considered. This process if executed in conjunction with specimen fabrication can provide feedback on the specimen preparation process.

Reliability is often characterized by a POD curve. POD and POFA should be integrated to generate ROC curves. Theory of signal detection methods allows estimates of bias in answering, independent of correctness, and estimates the true ability of the person to discriminate target situations. A usable basic technique is to ask inspectors about how certain they are of their identification of cracks.

A set of protocols is required for monitors and inspectors to assure consistent execution towards pre-defined goals.

Logistical planning is the first stage of implementing an experiment. The plan must ensure that facilities are suitable for the experiment, that they can provide the resources necessary, and that the experiment can be safely and efficiently executed. Typical steps involved in logistical planning are (1) assembly of hardware and dress rehearsal, (2) schedule of experimental sessions, (3) safety considerations, (4) storage and shipment of specimens, (5) field adjustments to loss or alteration of test specimens, and (6) post experiment archiving of test specimens. The implementation process consists of (1) preparation, (2) experiment execution, and (3) data qualification. Further details on various aspects can be seen in the report [Spencer 93].

A parallel report [Spencer 93a] details the procedures, hardware and techniques used to field an experiment to quantify the reliability of high-frequency eddy current inspections of aircraft lap splice joints. These two reports formed the basis for the ECIRE program discussed in details earlier.

1993 Protocols (FAA-Sandia)

The only report containing information on developing protocols was written as part of the FAA program with Sandia National Labs [Spencer 93]. Protocols are required to regulate complex experiments involving both experimental data gathering and human factor assessment. The primary functions of a protocol are to assure that: (1) objectives of the experiment are implemented, (2) consistent information is given to the inspectors and

their managers, (3) the experiment is carried out in a consistent manner, (4) recorded data are defined and gathered consistently, (5) deviations from the experimental plan are dealt with effectively, and (6) subsequent experiments can be carried out. The general areas of monitor protocols are: (1) operating the test equipment, (2) conducting briefings and questionnaires, (3) managing the tests, (4) providing test specimen quality assurance, (5) observing the inspector, (6) providing on the spot information, (7) controlling the documentation, (8) interacting with the inspector, (9) recording test conditions and environments, (10) recording inspection results, and (11) analyzing the data. The general areas of inspector protocols are (1) questionnaires, (2) structured briefings, and (3) inspection procedures.

1997 Field NDI Reliability Study Designs to Incorporate Human Factor Issues

Spencer discusses the process of designing a reliability study to quantify reliability associated with an NDI process, including the human operator [Spencer 97]. In planning for an assessment program, factors relating to the human operator and the environment in which the inspection takes place should be identified. Those factors that can be controlled cost effectively, should be included in a statistically based experimental design. Those factors that cannot be controlled or are prohibitively expensive should be given a chance to operate by locating experiments in the usual inspection environment and by simulating normal inspection conditions as close as possible. The experiment should not compromise the usual activities of those that are involved in the inspection process. That is, inspectors should not be asked to perform tasks or use equipment or procedures that are not a part of the inspection process that they would normally follow.

By identifying variables that potentially influence reliability and using binary regression models (e.g., probit and logistic), the experiment can be evaluated for its ability to provide clear information about those variables during the planning stages. The evaluation is accomplished through the application of traditional statistical techniques for determining estimability and confounding patterns.

Implementation of the experiment should be governed by specific protocols. The act of specifying a set of protocols and the subsequent following of those protocols helps ensure that data are consistently gathered from start to finish and from location to location. Binary (hit/miss) data can be augmented by soliciting ratings from inspectors concerning their calls. Points on a receiver operating characteristic (ROC) curve can be estimated by altering the criterion level used for treating the positive calls as hits. The result will be a better understanding of the trade-off between false call rates and detection rates.

In traditional models, POD(a) approaches 1.0 as the flaw size, a, gets large. These models should be generalized to include a parameter, c, that limits POD(a) to always being less than 1-c. The parameter c should be estimated from the data. The parameter, c, represents a miss rate that is independent of flaw size. Such a parameter is necessary when considering that many of the human factors that can affect an inspection would do so independent of flaw sizes.

13.3 From Other Programs

1996 Design of Statistical Methods in NDE

NDE reliability experiments are designed for two different purposes: assessing measurement capability and improving the measurement process. Measurement capability experiments attempt to study how both systematic and random variability are introduced into the measuring process by the various controlled and uncontrolled factors. Emphasis is placed on identifying the important factors that influence the POD and ROC curves. Measurement process improvement efforts try to find settings of controlled variables that simultaneously minimize the deleterious effects of uncontrolled variables and improve POD or ROC. Measurement capability studies provide a snap shot of the current measuring process. Process improvement studies are pro-active.

In an experiment, the factors can be fixed or random, crossed or nested. The various examples are: (1) fixed factors such as NDE technique, inspection material type, couplant type, instrument type, and scan rate; (2) random effects such as specimen, defect, facility, inspectors, and environment; (3) crossed factors such as inspectors and specimens, inspectors and NDE technique, facility and specimens, and scan increment and specimen; and (4) nested factors such as facility and inspectors, technique and inspectors [Olin 96].

1998 Design of the Experiment for NDE Capabilities Assessment

Issues in the *NDE Capabilities Handbook* [Matzkanin 98] are summarized in an NTIAC state-of-the-art report [Easter 98]. The critical factors in design of the experiment are: (1) selecting and producing cracks that are representative of NDE application and test object, (2) re-sampling the same crack leading to lack of crack-to-crack variation, (3) producing cracks at a size near 90% detection threshold, (4) using controlled NDE procedures for repeatability, and (5) designing data collection under variable conditions. The three methods for data analysis summarized are the modeling POD response ($\hat{a}$ vs. a), accept/reject data (log-odds model), and the point estimate methods.

1998 NORDTEST Guidelines for NDE Reliability

This NORDTEST technical report serves as a guideline on evaluation and tests to be performed when it is desirable on rational basis to determine NDE reliability, but to cover only the characterization of defects in materials [Forli 98]. According to NORDTEST, NDE process includes two main stages: defect detection (POD) and defect characterization (size, location, and type). POD should be measured against defect severity (size and type) to allow for a direct use of NDE results in structural integrity assessment. Estimates of POD values based on demonstration trials with a limited number of defects has an uncertainty, which must be taken into account, for instance, by using confidence levels or standard deviations. POFA, however, cannot be related to any specific quantity, like discrete defect sizes. One way is to relate it to unit of the object examined such as per meter or per foot or per element defined. When the sensitivity of

an NDE equipment is increased by adjusting response threshold parameters, both POD and POFA increases. This relationship can be shown in a ROC diagram.

Information must be provided for the confidence in determined NDE reliability quantities, for the quantities to be of any value. *Standard Confidence Criterion* is "the difference between average value and the lower 95% confidence limit of an estimated probability value for the NDE technique in question shall be less than or equivalent to 0.1." This criteria may be applied to relevant groups of defects or relevant points on POD curves, as well as to probability values on correct characterizations. *Standard Accuracy Criterion* is "the standard error in estimated (measured average) systematic deviations shall be less than 18.6% of the standard deviation of a single measurement."

The assessment program for reliability determination comprises (1) collection of available background material including technical description of the NDE technique and its performance; (2) initial evaluation and conclusions based on available information; (3) identification and evaluation of significant parameters and their variability; (4) planning and execution of NDE capability test program; (5) planning and execution of NDE reliability test program; (6) reference investigations; (7) evaluation of results from capability and reliability trials; and (8) calculation and presentation of reliability quantities.

The purpose of the capability test program is to demonstrate that the NDE technique is capable of detecting defects of intended types and sizes within the operational parameter windows, or to set limits to these. The capability test program typically includes a test matrix covering possible worst-case settings of the essential parameters and may be conducted on test samples with artificial defects. The aim of the reliability test program is to produce documentation on the reliability of the NDE technique by providing reliability quantities. The reliability test program may be conducted using laboratory-fabricated test specimens with natural defects, or relevant parts of real construction, relying on the fact that these contain a representative selection of naturally occurring defects. The number of defects and specimens shall be selected according to requirements of statistical confidence and ability to provide sufficient variability of essential parameters. During the testing process, sufficient measures of secrecy are required in order not to bias the NDE operator's prior knowledge. The tests shall be blind tests. As far as practically possible, the defects contained in the objects for the reliability tests should be individual ones at distances sufficiently far apart for the defects not to interfere.

To approximate POD curves, a form like the cumulative log-normal distribution or the log-odd model is suitable.

Modeling offers an economical alternative to experimental determination of NDE reliability quantities. This alternative involves a physical modeling of the NDE process and system as a function of all relevant parameters. Due to complexity of models and uncertainties in parameter values and distributions, a model might have to be subject to experimental verification. Models also play an important role in pinpointing important

parameters, which, if not sufficiently well known, can be subject to experimental verification.

13.4 Observations

- Best recommendations for design of NDI reliability experiments can be seen from FAA-sponsored work with Sandia National Labs [Spencer 93]. MIL-STD and other reports from Dayton Research Institute also contain very valuable information [MIL-STD-1823 (draft), Berens 88, and Hovey 89]. Some of the major concerns and issues that should be addressed by NDI reliability experiments are:
  - Methods of controlling the factors to be evaluated in the experiment
  - Methods of identifying, tracking, and accounting for the uncontrolled factors
  - Specimen design, flaw size range, and identification
  - Flaw characterization, and specimen maintenance
  - Human factors, particularly the awareness of being evaluated
  - Size of the experiment for a meaningful conclusion
  - Facility sampling and inspector sampling
  - NDI variables such as inspection equipment, materials, procedures, process control, calibration standards and methods.
  - Level of detail in monitor and inspector protocols for consistency without influencing the natural performance style
- The word capability and reliability should not be confused. Capability reflects potential to perform. Reliability reflects actual "probabilistic performance."
- The Log-odds model is the best known method for data analysis.
- POD and POFA are both significant.

Best Recommendations: FAA-Sandia

Good Recommendations: MIL-STD-1823 (Draft)[9]

[9] Or recently accepted MIL-HDBK-1823

14. RELIABILITY MODELING AND PREDICTION

Over the last decade, engineers have started making efforts to predict reliability through mathematical and computational models. Most of this work has come from Universities and R&D labs. Implementation of the damage tolerant design approach rests on three methodologies: stress analysis, NDI, and failure modeling. Extensive capabilities are in place for modeling stresses and failures, and are widely used in the design process. However, modeling of NDI is not nearly as widely accepted; instead frequent use is made of empirical rules based on extensive demonstration programs. For both economic and time reasons development of a model base for estimating NDI reliability is greatly needed. Integration of NDI reliability models with CAD systems can allow assessment of inspectability during the design stage [Gray 89]. Since the subject has potential to reduce future experimental reliability assessment efforts, it deserves a separate chapter for review.

14.1 Modeling and Prediction Efforts

1989 Models for Predicting NDE Reliability in Engine Components

Gray et al. present models for predicting NDE reliability based on certain characteristics [Gray 89]. First, the model must predict the response of a real measurement system, as influenced by the specific characteristics of commercially available probes and instruments, rather than an idealized response based on assumptions. Second, the models should give as outputs the information obtained by real protocols; if separate protocols are followed in detection and sizing, these should be described by separate models. Third, the models should be used to develop more reliable standardization approaches. Fourth, the models should be configured such that they can be integrated with standard CAD packages. The paper describes computer models that can be used to generate quantitative predictions of inspectability for ultrasonic, eddy current, and x-ray film methods. They are all physically based models that explicitly consider the characteristics of the realistic flaws, the typical component geometries and materials, and the practical inspection methods and practices.

1990 Computer Modeling of Eddy Current POD

Beissner and Graves demonstrated feasibility for the prediction of the probability of crack detection in eddy current inspection of complex geometry parts [Beissner 90]. The computer model used for this purpose consisted of a set of algorithms for predicting probe impedance and analyzing such data in terms of flaw detection probabilities. The algorithms include: (1) a three-dimensional boundary element code for the calculation of the magnetic scalar potential on the surface of a flawed or unflawed part, (2) an adaptation of the boundary element method to the prediction of the field produced by a

non-axis-symmetric probe with a ferrite core and shield, (3) a probe impedance code that uses output from the boundary element codes to determine changes in probe impedance as a function of probe position, and (4) programs for using calculated flaw signals and experimental noise data to predict the probability of flaw detection. The operation of this package of codes is described and illustrated by calculation of the probability of detection of cracks in a circular beveled edge of an aircraft structural member. However, at the end of the manuscript, the author says that considerably more effort will be required to develop tools that can be reliably used for eddy current POD estimation.

1990 Modeling Inspectability for an Automated EC Measurement System

Nakagawa et al. assembled a computer-controlled eddy-current workstation [Nakagawa 90a]. Experimental data from a series of measurements over a variety of flaws and materials with a uniform-field EC probe are presented to demonstrate its flaw detection capabilities. The quantitative NDI capability results for tight fatigue cracks are given in the form of ROC curves. Their model-based approach combines theoretical and experimental methods. They used theoretical calculations to determine the expected impedance signal due to a tight crack. They estimated the variability of impedance measurements from data accumulated in a series of calibration measurements. Those noise measurements were performed for several EDM notches of known size. The software developed consisted of system-control, signal processing, and POD packages. In another complimentary report [Nakagawa 90], authors demonstrated the capability of a computer simulation in replacing a large number of EC measurements to assess fatigue crack detectability.

1993 POD Models for ET

Rajesh et al. present POD models for ET [Rajesh 93, Rajesh 93a, Rajesh 93b]. They use a finite element measurement model to predict the measurement values. The factors influencing the measurement are perturbed to generate the ensemble of signals characterized by conditional probability density functions. The POD and POFA are then estimated by appropriate integration of the density function. Such models constitute a powerful tool for a wide variety of issues relating to NDE reliability. These models can provide insight into factors affecting detectability, assist in determining optimum test parameters, play a role as a vehicle for interpolating and extrapolating experimental POD results, and lead to savings in situations where complex defect shapes are difficult to replicate in a laboratory in large numbers.

1993 Model for Predicting Ultrasonic Pulse Echo POD

Ogilvy describes a mathematical model for predicting the theoretical POD of planar buried defects, for conventional ultrasonic pulse-echo inspection [Ogilvy 93]. A model for the scattering of ultrasound by well-oriented planar defects is combined with noise theory to produce a calculated capability of detection, based on the likelihood that the defect signal exceeds the specified threshold. The model also addresses the problem of

false indications (recording of a defect when none is present) by showing how any improvement in POD predictions must be considered in parallel with the associated change in the POFA. The author gave examples to illustrate how the model may be used to check proposed inspections. He also showed how factors such as reporting threshold, probe scan pattern, and criterion for the number of probe positions at which an indication must be seen before a defect is recorded affect the probability of detection and false indication. Ogilvy studied the effects of defect roughness on detection probabilities. He also used the model to quantify the uncertainties that result in POD predictions when defect properties such as orientation, roughness, aspect ratio, and depth within the specimen are themselves uncertain. The problems of equipment and human error were not addressed, although the discussion explained how, if these can be quantified, they may be incorporated into the model. His models were valid only for pulse-echo inspection of defects, which are detected through scattering and not diffraction.

1993 Uses of Model Based POD curves

Thompson and Schemer demonstrate the role that POD models can play in improving the capability and reliability of NDE measurements, in guaranteeing the reliability of components to which they are applied, and in designing new components with built-in inspectability requirements [Thompson 93]. They first describe the elements that go into formulation of the POD model using UT as a specific example. Then, they discuss how POD vs. flaw size curves obtained from such models can be used to define NDE system capabilities and determine the influence of inspections on component reliability. First, the models must be capable of simulating the fundamentals of the measurement processes to the point where actual measured signals received can be predicted, (i.e., complete measurement models must be available). Second, the important sources of variability in the measurement must be included and used to estimate the POD of defects. In an NDE experiment, these variability sources include: (1) operator, (2) defect character, (3) technique and procedure, (4) equipment, and (5) component geometry and surface property. Operator dependencies are still outside the curtain of calculability.

1996 Methodology for Estimating NDE Capability

Meeker et al. produced a significant methodology for combining physical modeling of an inspection process with lab and production data to estimate NDE capability. The proposed physical/statistical models could be used to predict POD, POFA, and ROC function curves. The general approach was to (1) develop the mathematical model for the physics of the inspection process, (2) use lab experiments on synthetic flaws to evaluate as many aspects of the flaws as possible, (3) revise the models to reduce the deviation, (4) apply the model to a set of production/field inspection conditions, and (5) revise the model. Their focus was on UT for detecting hard alpha and other subsurface flaws in titanium using gated peak detection. Iowa State University was conducting the experiments at the time of this report [Meeker 96]. FAA supported this work under the Engine Titanium Consortium program.

1998 Computational Modeling of POD

A number of computer models are now available for prediction of inspection reliability in terms of POD and false calls (PFI) [Wall 98]. These cover an increasing range of inspection techniques and run in real-time on a standard PC. The models are being increasingly used and validated. The computational modeling approach provides complimentary data to experimental assessments and allows existing experimental data to be more widely used. The models can further provide specific data not available from experimental measurements such as parametric studies, assessment of historical data, and optimization at the design stage. POD models are already being used in economic assessments, integrity assessments, to support safety cases, and validation of inspection procedures and plans. The values of POD now being obtained by modeling are not dissimilar in accuracy to those obtained in experimental trials. Methods are evolving to correct model predictions for human error.

The author (Wall) expressed that POD models should be seriously considered as an integral part of future POD trials. Models could reduce the number of test samples required, help gain acceptance and familiarity for the modeling approach, provide validation, and lead to improvements in model predictions and correction methods used for human and environmental effects.

14.2 Observations

- The subject of computational models for POD prediction is relatively recent; such modeling was probably waiting for computational tools to mature.
- Predictive models can complement and validate limited and expensive experiments, reducing the number of specimens required, lengthy tests, and overall cost of reliability assessment.
- Models can be integrated with CAD systems and damage growth models for optimization during design at the level of life cycle costs.
- Models can provide specific data not available from experimental measurements and low cost parametric studies. They can be of extensive help in extrapolating POD data obtained on simple specimens to more complex applications.
- Models must be validated by correlation to experimental data before being used in serious applications.
- Models can help apply corrections for factors that could not be controlled during the experiment and even some of the human factors.
- Predictive models need further research.
- Not much could be found on inspection simulation.

Nothing worthwhile identified

15. OTHER RELEVANT REPORTS

This chapter covers some of the other reports that are related to the subject under investigation, but did not fit very neatly into the chapters classified so far.

15.1 European American Workshop on NDE Reliability, 1997

NDE practitioners from Europe and USA assembled for a 3-day workshop in Berlin to come to a common understanding of NDE reliability and different ways for its determinations. The invited presentations focused on different technical and scientific approaches to the problem of how to guarantee or demonstrate NDE reliability and dissimilar qualification concepts used in different industries in Europe and North America.

During the workshop it became clear that no absolute truth exists on how to determine the reliability of NDE, especially on how to quantify the human factors. However, there appears to exist a variety of promising approaches and a valuable pool of experiences such as the Program for Inspection of Steel Components (PISC), Performance Demonstration Initiative (PDI), European Network for Inspection Qualification (ENIQ), Engine Titanium Consortium (ETC), and NORDTEST. Also it was agreed that much of the misunderstanding within the technical community can be traced to the use of different vocabularies or application of the same word to different things. Consequently creation of a common dictionary was identified as necessary. The participants agreed that no absolute reliability of NDE is available; instead every NDE system must be qualified and validated on a case-by-case basis.

The group proposed and agreed to a mathematical formula for NDE reliability put forward by Serge Crutzen and Matt Golis. This formula facilitates objective handling of the terms capability and reliability and supports a rational design of qualification/validation rules. It was stated as

$$\mathbf{R} = f(\mathbf{IC}) - g(\mathbf{AP}) - h(\mathbf{HF})$$

R, Reliability of an NDE system applied or performance of the procedure application;

IC, Intrinsic Capability of the system generally considered as an upper bound;

AP, effect of Application Parameters, such as access restrictions and surface state; generally reducing the capability of an NDE system.

HF, effect of Human Factors, generally reducing the capability or effectiveness.

This model gave an acceptable frame for discussion for both US and EU aeronautical, nuclear, and other industries. It offers a good way of identifying which of the elements is considered during round-robin tests, parametric studies, and data discussions.

The issues with organization of practical trials were discussed. One of the limitations is that they only form a part of the process of qualification/performance demonstration. Also needed are quality assurance and technical justification. The defect characterization is based on parameters such as flaw type, orientation, number, and sizes. Detection criterion can vary greatly depending on the situation. Reality of specimens that mock the inspection varies a lot; some type of a protocol could help. Depending on technical need, it is possible to use either or both open trials and blind trials. Another conclusion was the need to establish correlation between artificial and real defect characteristics, either empirically or theoretically through modeling. Modeling can be used to assess POD when experimental studies are impractical. Simulators can be used to provide input to statistical studies regarding human response. However, validation of such models remains to be a central issue.

15.2 American European Workshop on NDE Reliability, 1999

The second joint workshop was conducted at Boulder to debate the reliability model proposed in the first workshop and discuss issues related to development and study of individual parameters.

Metric to Measure Performance

NDE reliability should be defined as a combination of high POD for the defect sizes of concern and a low POFA for small non-existent defects. If possible, a characterization of the distribution of the defect sizes should be considered in the definition of inspection reliability. If the NDE reliability is to be separated into IC, AP, and HF, agreement must be reached regarding variables that comprise the factors for each basic inspection method. These variables will determine if a deterministic or stochastic characterization is required. A model is also needed to combine the contributions of the individual factors and their interactions to POD and POFA for the system [Berens 99].

From the European aerospace viewpoint, the standard is >90% POD with > 95% confidence and <3% POFA [Tober 99].

Concern over the Proposed Model

Taylor and Doctor expressed concern on the proposed mathematical model [Taylor 99]. They believe that the process of defining the functions f(IC), g(AP) and h(HF) and developing a mathematical relationship that would enable one to quantify NDE reliability could be very costly and can take many years to develop a practical methodology. They recommend proven reliability assessment methodology that requires these basic steps: (1) define NDE as a system, (2) define the required NDE system performance, (3) define a technically defensible measure of achieved performance, and (4) develop a technically defensible correlation between achieved performance and required performance. In their viewpoint an NDE system contains two components - data acquisition and data

evaluation, and it is easier to develop a formulation that may be assessed by conventional reliability methodology. Once the NDI is defined as a system, the interdependence of system components can be defined, and reliabilities of each component can be combined using accepted mathematics. The process of defining NDE system components should provide a level of detail that provides an accurate assessment of reliability, within limits of cost effectiveness.

Tober points out that the inspector should not be disregarded as a factor, as he is capable of compensating for the shortcomings of the inspection procedure or equipment [Tober 99]. Further discussion is required on whether all process steps are contained in the proposed formula and if they are appropriately assigned to the three terms.

Definitions

The word **IC** was redefined from Intrinsic Capability to Ideal Capability. Few other definitions were arrived at. They are quoted as presented in 1999 workshop.

NDE System is the procedures, equipment and personnel that are used in performing NDE inspection.

Reliability - NDE reliability is the degree that an NDT system is capable of achieving its purpose regarding detection, characterization and false calls.

Ideal Capability is the hypothetical optimal performance of an NDE technique based on the governing physical principles.

Application Capability is the degree, to which an NDE system achieves its intended purpose, excluding human factors. It is defined in the context of the specification of expected application parameters.

Application Parameters are the factors concerning material conditions, discontinuities, procedure and equipment that influence the ability of an NDE system to meet its intended purpose.

Detection - Threshold-driven identification of the existence of a signal/indication to be of interest or worthy of further investigating.

Signal/Data Interpretation - Deciding relevance of a signal indication as being valid for further indication/materials characterization. e.g. geometrical reflection vs. cracks.

Indication Characterization - Estimation of size, location, orientation, type, nearest neighbors.

Indication Evaluation - Comparing characterizations to acceptance criteria for the purpose of making accept/reject decisions.

15.3 General Related Reports

1976 NDI Measurements: How good are they?

In ASNT's Mehl Honor Lecture titled "Non-destructive measurements: How good are they?" Harold Berger stressed: "If we are to have measurement assurance in an NDT system, each part of the system has to have a known level of accuracy. We need data on the adequacy of measurements" [Berger 76]. However collecting accuracy data is difficult without giving due consideration to false calls [Lewis 78].

1978 Determination of NDI Reliability using Field or Production Data

Most of the major programs on NDI reliability assessment have used specimen testing as the core of the program. Johnson had fairly early introduced a method that used field or production data rather than specimen data to estimate the probability of rejecting a part containing an imperfection of a given size [Johnson 78]. The NDI signal characteristics were correlated to the relative size of the imperfection from destructive examination of the material units rejected during field examination. The advantages of this method were cost savings and actual NDI on imperfections rather than artificially induced flaws. However, literature shows that this method did not become popular. This is in large part due to missing data on flaws that are missed in field and production applications.

1992 POD - GE Aircraft Engines Experience

Domas expressed a genuine practical concern about the Engine Structural Integrity Program (EnSIP) POD data [Domas 92]. Extensive, valuable EnSIP POD data in large quantities have been generated demonstrating the importance of quantifying NDE and leading to significant systems development and improvements. The limited data in that lab specimens can only quantify capability and not specific crack detectability. "Ultimate POD" is a function of both the NDE system and the implementation environment in the broadest terms. "Real" parts and environments must be addressed in order to confidently employ risk management strategies and develop reliable logistics planning. POD technologies such as an ROC curve (or other alternatives) need to be pursued with the same vigor and coincident with NDE technique development.

1995 Aging Aircraft NDI Validation Center

The National Aging Aircraft Research Program has established the FAA Aging Aircraft NDI Validation Center at Sandia National Labs in Albuquerque, NM. The center is accessible to airline operators, manufacturers, equipment vendors, and other interested organizations wishing to test or observe the testing of current, enhanced, or emerging maintenance and inspection techniques, equipment, and systems. POD studies are a corner stone of the center. The ECIR experiment was a tremendous success [Walter 95]. The center is indeed a very valuable resource.

1997 NDE Capabilities Data Book

NDE Capabilities Data Book prepared and regularly published by NTIAC [Matzkanin 97, Rummel 96], consolidates and organizes available reference data for demonstrated NDE performance capabilities into a single source. Guidelines are presented for selecting options for use of NDE and for assessing the potential to meet design requirements in terms of critical flaw detection. Guidelines for demonstration of specific NDE process capabilities are also presented. Following an exhaustive text description, over 400 POD curves are organized by NDE method. POD data are generally presented as a function of crack length and, at times, crack depth and crack depth-to-thickness ratios. Original reference source information is provided for each set. Procedures covered are ET, UT, RT, PT, MT, Visual, and other emerging processes. Materials covered are AL2219 T-87, AL 2024 T-37, AMS 355 Steel, and Ti 6Al4V. This work is available in both hard copy and in CD form. The CD form contains raw data supporting each POD curve presented and is intended for use in assessing and validating new analysis and modeling methods.

1997 and 1999 Overview of NDE Capability and Reliability

Ward Rummel provides an overview of NDE capabilities in the aerospace industry [Rummel 97] and NDE reliability for aging aircraft [Rummel 99]. He listed the myths and misunderstandings about NDE, such as: (1) NDE is absolute and no flaws exist after inspection, (2) detection and discrimination capabilities are at the calibration level, (3) increase in amplifier gain increases the detection level, (4) NDE response from slots used for calibration is the same as the response from a crack of equal size, (5) all cracks of the same size are created equal, (6) cracks are equally detectable under lab/factory and field conditions, (7) critical crack size criteria are applicable everywhere on a part, (8) all certified NDE personnel perform at the same level, (9) an NDE measured flaw is the actual flaw size, and (10) optimized NDE procedure is the one that detects the smallest crack.

Rummel summarizes the past POD efforts as follows. POD as a method of assessing and quantifying NDE capability originated in the US aerospace industry, and a large amount of available data have been generated. These data are primarily from fatigue cracks in flat plate specimens. Transfer of flat plate data to more complex shapes may be accomplished using equivalent response and equivalent signal/noise response methods. Slight procedural and personnel performance variances may be assessed using subsets of full qualification/validation of test artifacts. Focus on calibration and casual model methods offers the best approach for assessing and quantifying reproducibility of a specific NDE procedure. Full end to end POD assessment is necessary for new applications.

Reliability of NDE procedure may be characterized in terms of detection capability (POD), reproducibility (calibration), and repeatability (equipment, material, process, and procedure). The capability of the procedure is inherent to the NDE mode and method; reproducibility and repeatability are application controllable parameters, which can be

used to predict and validate the performance reliability of a given procedure. Variations in reproducibility and reliability are equally applicable to automated scanning systems as they are to procedures involving manual scanning.

15.4 Observations

- *NDE Capabilities Data Book* is a valuable source of information
- FAA Sample Defect Library at the Aging Aircraft NDI Center (AANC) at Sandia Labs is another valuable source of specimens. The library contains small sized specimens as well as full-scale actual aircraft specimens.
- During the European-American workshop on NDE reliability, it became clear that no absolute truth exists on how to determine the reliability of NDE, especially on how to quantify the human factors.
- The workshop highlighted vast differences in NDI procedures, test objects and requirements in various industries. One of the great difficulties with the workshop was in identifying meaningful common factors among the various industries. The resultant simplified formula is useful for purposes of discussion, but has little relevance to the task of assessing useful detection capabilities in any industry.
- Participants at the workshop also agreed that no absolute reliability of NDE exists, rather every NDE system must be qualified and validated on a case-by-case basis.
- There is value in determining what actual performance is against what best can be achieved.

Useful Events: Joint American European Workshop

16. REMARKS

16.1 Summary of Observations

Programs

- The subject of NDI capability and reliability assessment started drawing serious engineering attention in early 1970s with the Space Shuttle and B-1 programs. Initial field assessment programs with Air Force on aircraft and engine components had revealed deficiencies in detection reliability. Subsequent efforts, including training and re-evaluations have improved technician proficiency. Continuous reliability assessment and improvement appear imperative to maintain force readiness.
- Different programs have been supported by different organizations/industries with slightly varying objectives. However, most programs have been philosophically similar – assessment of reliability. Programs do not appear to evaluate what best could have been achieved; in some sense, the ideal capability measurement did not receive proper attention. Recent programs do not show clear evidence of using experience from earlier programs.
- AF-Battelle, AF-SwRI, and development of MIL-STD programs offered AF with guidelines and tools for sustained NDI reliability assessment. However, no evidence of their continuous use could be seen. Recommendations from the MIL-HDBK-1823 can certainly be used in future programs.

Sampling

- To obtain an assessment of overall NDI capability, an average of all inspections that are conducted by the AF is required, which can be cost prohibitive. Facility and inspector sampling gains significance in this situation. Judgment-based selection of facilities and inspectors might provide more accurate data as compared to a totally random selection.

Specimens

- Specimen number and design, fabrication, flaw induction, flaw characterization, and maintenance hold the key to success of the program. Generally, specimens have been artificially fabricated specifically for the purpose of assessment. Some programs had a combination of synthetic and actual flawed parts.
- Environmental degradation of specimens has been out of the scope of most programs.
- As the statistical tools evolved, the specimen requirement reduced. The concept of racks with interchangeable subassemblies is a good way to eliminate prior knowledge of defects from inspectors who have already gone through the program.

- Specimens, data, and reports from AF-sponsored programs could have been better preserved, preferably in a single office with proper documentation.

Inspections

- Most of the programs involved testing in actual environment with the actual equipment that is used in production lines. These factors are very important for obtaining a realistic reliability estimate.
- Briefings to management and participating inspectors are an important part of the actual inspection section of the program. Inspectors need to be assured that their identity will not be disclosed and their performance will only lead to a data point in the analysis. Still, human factor specialists say that inspector's performance deviates from the everyday performance.
- Inspection scheduling need to address issues that are local to the facility such as overtime, shift work, and accelerated inspections.

Data Acquisition and Analysis

- Different programs had different ways to record defect data. Initial programs only concentrated on finds and misses, but later efforts recorded false calls and truly identified flawless specimens, as well as the level of confidence in the inspection result. Observation of facility characteristics is an important segment as the environmental parameters have bearing on the inspection.
- Statistical methods govern the design and magnitude of the experiment. The best of the statistical procedures for data analysis today are the log-odds model and the $\hat{a}$ vs. a method. The former is good for hit/miss type of data acquisition and later one is better supported by continuous-function type of NDI response. Different statistical procedures are likely to provide differing POD curves. These differences pose a challenge in trying to present the results at the end of tiring assessment experiments.
- A POD curve projects the NDI reliability in terms of detection only. This is important from a safety viewpoint. An ROC curve projects the NDI reliability in terms of both detection and false calls. This is important from a safety as well as an economic viewpoint. The CC is another parameter that provides a measure of individual reliability in reference to pure chance. It accounts for finds, misses, false calls, and true no-calls.

Human Factors

- Human factors that influence an inspector's discrimination and decision-making abilities can be classified into three categories: physical environment, organizational climate, and mental state. Physical environment involves temperature, humidity, illumination, noise, and posture. Most of these factors are somewhat controllable. Organizational climate involves management attitude, schedule pressures, personnel behavior, hierarchical structure, and wages. These are also somewhat controllable. Mental state deals with intrinsic attitude and interest in the subject of NDI, internal

motivation and zeal to excel, and daily fluctuating factors such as mood, fatigue, sleeplessness, monotony, family, and financial situations. These are really uncontrollable.

- Many investigations believe that if NDI engineering is not well defined, the human should not be blamed for poor performance. There continues to be a lack of good human factors investigation and mitigation programs.

Role of Computers

- Most programs have been dominated by field experiments. There is room and value for computer simulations and modeling to supplement the experimental data. The computational predictive models for POD can complement and validate limited and expensive experiments, by reducing the number of specimens required, lengthy tests, and overall cost of reliability assessment. Models can further provide specific data not available from experimental measurements and low-cost parametric studies.
- The most useful near term use of models may be in providing correction factors to account for variances introduced by differences in "calibration artifacts" and procedures; equipment variances; and in extrapolation of data from one material to another and from simple test coupons to complex shapes.
- Modeling procedures appear to be at an advanced stage of development. Inspection simulation tools are not so common even in R&D establishments.

Workshop Wisdom

- Participants of the joint European and American workshops on NDI reliability felt that no absolute truth exists on how to determine NDI reliability, and how to quantify the human factors. They also agreed that there is no absolute reliability of NDI; rather every NDI system must be qualified and validated on a case-by-case basis.

16.2 Useful Documents and Valuable Resources

- Guide to design and implementation of NDI reliability experiments [Spencer 93].
- Recommendations on NDI System Reliability Assessment [MIL-HDBK-1823].
- Account of NDI reliability data analysis [Berens 88].
- Single comprehensive document on NDT, reliability assessment, data analysis, and presentation with the discussions from peers [Olin 96].
- Compendium of POD curves - *NDE Capabilities Data Book* [Matzkanin 97].
- Hardware resource - Sample Defect Library at Sandia National Labs [Roach 95].
- Analysis software - recent version of POD/SS.

17. VISION FOR THE FUTURE

NDI reliability and capability assessments are required to be a continuous activity. From time to time, organizations also require special studies to assess the impact of various controlled and uncontrolled variables, equipment qualification prior to introduction into service, and other assessments. There will be an ever-increasing role of computers, in terms of databases and information management, inspection response modeling, inspection procedure simulations, decision support tools, performance analysis, prediction algorithms, and web-based activity.

Figure 17.1 shows a possible approach to NDI performance assessment, which can synthesize input through multiple modes. Once set-up, such a system can continuously provide considerable insight into NDI inspection activity and provide all the information required to understand existing performance and make changes as necessary to accomplish the NDI mission.

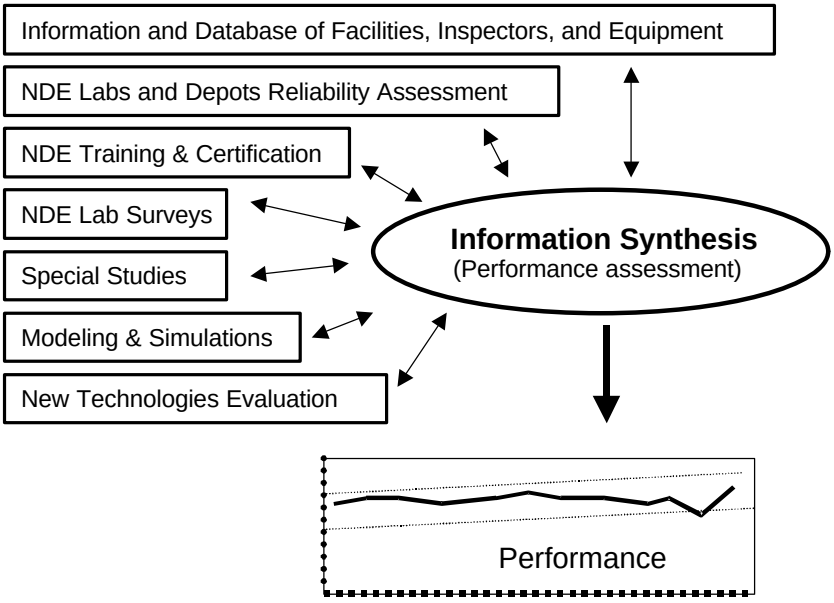

Figure 17.1: A vision of a future NDI performance assessment program. *Concurrent input from multiple modes will help provide complete insight into the organization-wide NDE reliability and capability, acceptable and desirable levels of performance, and long-term trends and short-term indicators. Comprehensive information database leads to better management insight and decisions.*

BIBLIOGRAPHY

[Ainsworth 85] L. Ainsworth, *Human Factors Considerations in NDI,* Proc. of 11th WCNDT, pp. 1115-1119, 1985.

[Anderson 73] R. T. Anderson, T. J. Delacy, and R. C. Stewart, *Detection of Fatigue Cracks ~ Nondestructive Testing Methods,* NASA CR-128946, March 1973.

[Annis 89a] C. Annis and K. Erland, *Measuring Differences among POD Curves,* Review of Progress in Quantitative NDE, Vol. 8B, pp. 2229-2233, 1989.

[Annis 97] C. Annis, *Role of Statistical Design in Measuring NDE Reliability,* Proc. of European American Workshop on Determination of Reliability and Validation Methods on NDE, pp. 265-272, 1997.

[Ashbaugh 95] M. Ashbaugh and F. W. Spencer, *Protocol Assessment of a Field Reliability Experiment at the NDI Validation Center,* Materials Evaluation, Vol. 51, pp. 827-828, 1995.

[Bahravesh 89] M. H. Bahravesh, S. S. Karimi, and M. E. Ford, *Human Factors Affecting the Performance of Inspection Personnel in Nuclear Power Plants,* Review of Progress in Quantitative NDE, Vol. 8B, pp. 2235-2242, 1989.

[Beissner 90] R. E. Beissner and J. S. Graves III, *Computer Modeling of ET POD,* Review of Progress in Quantitative NDE, Vol. 9, pp. 885-891, 1990.

[Berens 81] A. P. Berens and P.W. Hovey, *Evaluation of NDE Reliability Characterization,* AFWAL-TR-81-4160, Vol. 1, Dec 1981.

[Berens 82] A. P. Berens and P.W. Hovey, *The Effect of Inspection Uncertainty on Crack Growth Based Maintenance Scheduling,* Summary Paper, Prepared under Contract F33615-80C-5l40, February 1982**.**

[Berens 83] A. P. Berens and P.W. Hovey, *Statistical Methods for Estimating Crack Detection Probabilities,* ASTM STP 798: Probabilistic Fracture Mechanics and Fatigue Methods, pp. 79-94, 1983.

[Berens 84] A. P. Berens and P.W. Hovey, *Flaw Detection Reliability Criteria,* AFWAL-TR-84-4022, 1984.

[Berens 84a] A. P. Berens and P.W. Hovey, *Flaw Detection Reliability Criteria Vol. I: Methods and Results,* AFWAL-TR-83-4089, Vol. 1, 1984.

[Berens 88] A. P. Berens, *NDE Reliability Data Analysis,* Metals Handbook, Vol. 17, 9e, NDE and QC, ASM International, pp. 689-701, 1988.

[Berens 88a] A. P. Berens, P.W. Hovey, R. M. Donahue, and W. N. Craport, *User's Manual for POD/SS,* UDR-TR-88-12, 1988.

[Berens 97] A. P. Berens, *"a versus a" Approach to POD;* Proc of European American Workshop on Determination of Reliability and validation Methods on NDE, pp. 99-106, 1997.

[Berens 99] A. P. Berens, *What Metric Should be Used to Measure Intrinsic Capability ? Is it only POD and FCP - or is Theoretical,* Second American-European Workshop on NDE Reliability, Boulder, CO 1999, 1999.

[Berger 76] H. Berger, *NDI Measurements: How Good are They?,* Materials Evaluation, 1976, Vol. 34, pp. 18A-34A, 1976.

[Burkel 96] R. H. Burkel, D. J. Sturges, W. T. Tucker, and R. S. Gilmore, *POD for Applied UT Inspections,* Review of Progress in Quantitative NDE, Vol. 15B, pp. 1991-1998, 1996.

[Burkhardt 99] G. L. Burkhardt, J. S. Stotle, J. L. Fischer, J. N. Buckingham, M. D. Paulk, and M. J. Litman, *Reliability Study of MOI Imaging Inspection of C-5 Aircraft Fuselage, Phase II,* The Third Joint FAA/DoD/NASA Conference on Aging Aircraft, 1999.

[Bush 83] S. H. Bush, *Reliability of NDE,* NUREG/CR-3310, Vol. 1, 1983.

[Chang 76] F. H. Chang, J. R. Bell, T. C. Walker, J. M. Norton, P. F. Packman, and L. O. Gilstrap Jr, *Methods for Determination of the Sensitivity of NDE Techniques,* AFML-TR-76-246, Feb 1977.

[Chern 94] J. E. Chern, N. J. Yang, and H. P. Chu, *POD of Delaminations in Composites,* Int. Advances in NDT, Vol. 17, pp. 97-115, 1994.

[Christner 83] B. K. Christner and W. D. Rummel, *NDE Detectability of Fatigue-Type Cracks in High-Strength Alloys,* MCR-83-568, Martin Marietta, July 1983.

[Christner 88] B. K. Christner, D. L. Long, and W. D. Rummel, *NDE Detectability of Fatigue –Type Cracks in High-Strength Alloys – NDE Reliability Assessments,* MCR-88-1044, Martin Marietta, September 1988.

[Corbly 70] D. M. Corbly, P. F. Packman, and H. S. Pearson, *The Accuracy and Precision of Ultrasonic Shear Wave Flaw Measurement,* Materials Evaluation, Vol. 30, No. 5, pp. 104-110, May 1970.

[Davis 87] M. Davis and P. Aguilar, *Proficiency Evaluation of USAF SA-ALC Structural Assessment Testing,* SA-ALC/MAQCN, Jun 1987.

[Davis 88] M. K. Davis, *Proficiency Evaluation of NDE Personnel Utilizing the Ultrasonic Method,* Review of Progress in Quantitative NDE, Vol. 7B, pp. 1777-1789, 1988.

[DeNale 89] R. DeNale and C. A. Lebowitz, *A Comparison of UT and RT for Weld Inspection,* Review of Progress in Quantitative NDE, Vol. 8b, pp. 2003-2010, 1989.

[DeNale 90] R. DeNale and C. A. Lebowitz, *Detection and Disposition Reliability of UT and RT for Weld Inspection*, Review of Progress in Quantitative NDE, Vol. 9b, pp. 1371-1378, 1990.

[Domas 92] P. A. Domas, *POD - Some Aircraft Engine Experience Potentially Relevant to Aging Aircraft*, Proc. of the Int. Workshop on Structural Integrity of Aging Airplanes, pp. 160-165, 1992.

[Easter 98] J. K. Easter and G. A. Matzkanin, *NDE of Cracks in Aircraft*, NTIAC SR-98-04, 1998.

[Fahr 95] A. Fahr, D. Forsyth, M Bullock, and W. Wallace, *POD Assessment of NDI Procedures - Results of a Round Robin Test*, Review of Progress in Quantitative NDE, Vol. 14B, pp. 2391-2398, 1995.

[Fahr 98] A. Fahr and D. S. Forsyth, *POD Assessment Using Real Aircraft Engine Components*, Review of Progress in Quantitative NDE, Vol. 17, pp. 2005-2012, 1998.

[Fischer 98] J. L. Fischer, G. L. Burkhardt, J. S. Stolte, J. P. Buckingham, P. C. McKeighan and J Fitzgerald, *Reliability Study of Magneto-optic Imaging Inspection of C-5 Aircraft Fuselage*, NASA/CP-1999-208982, pp. 230-239, 1998.

[Forli 98] O. Forli and Group, *Guidelines for NDE Reliability Determination and Description*, NT Tech Report 394, Approved 1998-04, 1998.

[Forli 99] O. Forli, K. O. Ranold, and Group, *Guidelines for Development of NDE Acceptance Criteria*, NT Tech Report 427, Draft 1999-09-15, 1999.

[Glasch 87] K. J. Glasch, *Human Reliability in NDE*, Materials Evaluation, Vol. 45, pp. 907, 1987.

[Goodlin 94] D. L. Goodlin, *Airframe Inspection Reliability and Capability Assessment Program*, Final project Report 17-3836 for SA-ALC/LDN, Apr 1994.

[Gray 89] J. N. Gray, T. A. Gray, N. Nakagawa, and R. B. Thompson, *Models for Predicting NDE Reliability*, Metals Handbook, Vol. 17, 9e, NDE and QC, ASM International, pp. 702-715, 1989.

[Heida 89] J. H. Heida, *Characterization of Inspection Performance*, Proc of 12 WCNDT, Vol. 2, pp. 1711-1716, 1989.

[Herr 74] J. C. Herr, *Human Factors in NDE*, Second ASM Materials Design Forum on Prevention, pp. 226-241, 1974.

[Hoppe 98] W. C. Hoppe, *Hidden Corrosion Detection Technology Assessment*, The Second Joint NASA/FAA/DoD/ Conference on Aging Aircraft, NASA CP-1999-208982/Part1, pp. 349-358, 1998.

[Hovey 88] P. W. Hovey and A. P. Berens, *Statistical Evaluation of NDE Reliability in Aerospace Industry*, Review of Progress in Quantitative NDE, Vol. 7B, pp. 1761-1768, 1988.

[Hovey 89] P. W. Hovey, W. H. Sproat, and P. Schattle, *The Test Plan for the Next AF NDI Capability and Reliability Assessment Program,* Review of Progress in Quantitative NDE, Vol. 8B, pp. 2213-2220, 1989.

[Howard 95] M. A. Howard and G. O. Mitchell, *NDI for Hidden Corrosion in USAF Aircraft Lap Joints; Test and Evaluation of Inspection Procedures,* ASME: Structural Integrity in Aging Aircraft, AD-Vol 47, pp. 195-212, 1995.

[Hyatt 88] R. W. Hyatt, *Surveillance and Control of AF NDI Labs and Shops. Part-I Phase -I (subtitle unknown),* No report number, 1988.

[Hyatt 88a] R. W. Hyatt, *Surveillance and Control of AF NDI Labs and Shops. Part-II Phase -I (Development of NDI Process Assurance Methodology for Field Lab Operations),* No report Number, Feb 1988.

[Hyatt 89] R. W. Hyatt, G. E. Ketcher, and R. G. Menton, *Surveillance and Control of AF NDI Labs and Shops. Part-I Phase -II -- (Validation of NDI process assurance methodology for engine O/H depots),* No report Number, Sept 1989.

[Hyatt 89a] R. Hyatt, *Surveillance and Control of AF NDI Labs and Shops. Part-II Phase -II -- (subtitle unknown),* No report Number, 1989.

[Hyatt 91] R. W. Hyatt, G. E. Kechter, and R. G. Menton, *POD Estimation for Data Sets with Rogue Points,* Materials Evaluation, Vol. 49, pp. 1402-1408, Nov 1991.

[Johnson 78] D. P. Johnson, *Determination of NDI Reliability Using Field or Production Data,* Materials Evaluation, Vol. 36, No. 1, pp. 78-84, Jan 1978.

[Karimi 87] S. Karimi, *Human Factors Affecting NDE Technician Performance,* Electric Power Research Institute, RP1570-19, 1987.

[Knorr 74] E. Knorr, *Reliability of Detection of Flaws and the Determination of Flaw Size,* AGARD Fracture Mechanics Survey, AGARD-AG-176, pp. 396-413, Jan 1974.

[Lewis 74] W. H. Lewis and W. H. Sproat, *Reliability of NDI on Aircraft Structures,* Interim Report Contract No. F41608-73-D-2850, P0038, Aug 1974.

[Lewis 75] W. H. Lewis, W. H. Sproat, and B. D. Dodd, *Reliability of NDI on Aircraft Structures - Test Plan,* Interim Report Contract No. F41608-…., Jul 1975.

[Lewis 78] W. H. Lewis, B. D. Dodd, W. H. Sproat, and J. M. Hamilton, *Reliability of NDI - Final Report,* Report No. SA-ALC/MME 76-6-38-1, Dec 1978.

[Lewis 78a] W. H. Lewis, W. M. Pless, and W. H. Sproat, *Govt. /Industry Workshop on the Reliability of NDI,* Workshop Proceedings, SA-ALC/MME 76-6-38-2, Aug 1978.

[Lewis 80] W. H. Lewis, W. H. Sproat, and B. W. Boisvert, *NDI Technician Proficiency Evaluation (ET, UT) - Test Plan,* Report from Contract No. F41608-79-D-A012, Nov 1980.

[Lewis 80a] B. W. Boisvert, W. H. Lewis, and W. H. Sproat, *AF NDI Personnel Certification,* Workshop Proceedings, LG82ER0098, Mar 1982.

[Lewis 81] W. H. Lewis, W. H. Sproat, and J. M. Hamilton, *AF NDI Technician Proficiency,* Final report, AG82ER0099, 1981.

[Lord 74] R. J. Lord, *Evaluation of the Reliability and Sensitivity of NDI Methods for Ti Alloys,* AFML-TR-703-107, Jun 1974.

[Lovejoy 95] D. J. LoveJoy, *A Practical Approach to POD with MT,* Insight, Vol. 37, No. 12, pp. 974-977, 1995.

[Malpani 76] J. K. Malpani and P. F. Packman, *On the Applicability of Fracture Mechanics - NDI Techniques to Critical Aircraft Structures and Disc Components,* AFOSR-TR-89-0678, 1976.

[Matzkanin 97] G. A. Matzkanin and W. R. Rummel, *NDE Capabilities Data Book,* NTIAC-DB-97-02, 1997.

[Matzkanin 98] G. A. Matzkanin and J. K. Easter, *NDE of Hidden Corrosion,* NTIAC SR-98-03, 1998.

[Meeker 96] W. Q. Meeker, R. B. Thompson, C. P. Chiou, S. L. Jeng, and W. T Tucker, *Methodology for NDE Capability,* Review of Progress in Quantitative NDE, Vol. 15B, pp. 1983-1990, 1996.

[MIL-A-83444] *Airplane Damage Tolerance Requirements,* MIL-A-83444, Jul 1974.

[MIL-I-6870E] *Inspection Program Requirements, NDI for Aircraft and Missile Materials and Parts,* MIL-I-6870E, 1979.

[MIL-STD-1530A] *Aircraft Structural Integrity Program, Airplane Requirements,* MIL-STD-1530A(11), 1975.

[MIL-STD-1823] Proposed MIL-STD for USAF: *NDE System Reliability Assessment,* (Draft) MIL-STD-1823, Aug 1989.

[MIL-HDBK-1823] Accepted form of MIL-STD-1823: *NDE System Reliability Assessment,* April 1999.

[Mordfin 80] L. Mordfin, *Reliability of NDE,* Critical Issues in Materials and Mechanical Engg, Vol. 47, pp. 133-147, 1980.

[Nakagawa 90] N. Nakagawa and E. Beissner, *Probability of Tight Crack Detection via ET,* Review of Progress in Quantitative NDE, Vol. 9A, pp. 893-899, 1990.

[Nakagawa 90a] N. Nakagawa, M. W. Kubovich, and J. C. Moulder, *Modeling Inspectability for an Automated ET,* Review of Progress in Quantitative NDE, Vol. 9A, pp. 1065-1072, 1990.

[Nockemann 99] C. Nockemann, G. R. Tillack, C. Bellon, and M. Scharmach, *What Metric Should be Used to Measure the Effect of Application Parameter and How*

should it be treated using a Modular Approach ?, Second American-European Workshop on NDE Reliability, Boulder, CO, 1999.

[Ogilvy 93] J. A. Ogilvy, *Model for Predicting UT Pulse Echo POD,* NDT&E Int., Vol. 26, No. 1, pp. 19-29, 1993.

[Olin 96] B. D. Olin and W. Q. Meeker, *Applications of Statistical Methods to NDE,* Technometrics, Vol. 38, No. 2, pp. 95-130, 1996.

[Otterloo 99] D. V. Otterloo, J. Miller, and D. Pettit, *Assessment of Notch Detection Capabilities in Typical Aerospace Structures,* The Third Joint FAA/DoD/NASA Conference on Aging Aircraft, 1999.

[Packman 68] P. F. Packman, H. S. Pearson, G. B. Marchese, and J. S. Owens, *The Applicability of a Fracture Mechanics -NDT Design Criterion,* AFML-TR-68-32, 1968.

[Packman 76] P. F. Packman, J. K. Malpani, and F. M. Wells, *Probability of Flaw Detection for use in Fracture Control Plans,* AFOSR-TR-76-0290, 1976.

[Petrin 93] C. Petrin, C. A. Annis, and S. I. Vukelich, *A Recommended Methodology for Quantifying NDI/NDE Based on Aircraft Engine Experience,* AGARD-LS 190, 1993.

[Petru 85] J. A. Petru, *U. S. Air Force Reliability Programs,* 11th Word Conference on NDT, Proceedings, Vol. 2, pp. 1120-1123, 1985.

[Rain 84] C. Rain, *Uncovering Hidden Flaws,* High Technology, pp. 49-55, 1984.

[Rajesh 93] S. N. Rajesh, L. Udpa, and S. S. Udpa, *Estimation of Eddy Current POD using FEM,* Review of Progress in Quantitative NDE, Vol. 12B, pp. 2365-2372, 1993.

[Rajesh 93a] S. N. Rajesh, L. Udpa, and S. S. Udpa, *Numerical Model Based Approach for Estimating POD in NDE Applications,* IEEE Transactions on Magnetics, Vol. 29, No. 2, pp. 1857-1858, 1993.

[Rajesh 93b] S. N. Rajesh, L. Udpa, and S. S. Udpa, N Nakagawa, *POD Models for Eddy Current NDE Methods,* Preprint copy obtained from author, 1993.

[Roach 95] D. Roach, K. Harmon, C. Jones, and P. Walkington, *Aircraft Inspection Validation Experiments and the Use of NDI Validation Samples,* Materials Evaluation, Vol. 51, pp. 803-807, 1995.

[Rudlin 92] J. R. Rudlin and L. C. Wolstenholme, *Development of Statistical POD Models Using Actual Trial Inspection Data,* British J of NDT, Vol. 34, No. 12, pp. 583-589, Dec 1992.

[Rummel 74] W. D. Rummel, P. H. Todd, S. A. Frecska, and R. A. Rathke, *The Detection of Fatigue Cracks by NDT Methods,* NASA CR 2369, Feb 1974.

[Rummel 75] W. D. Rummel, *The Detection of Tightly Closed Flaws by NDT Methods,* MCR-75-212, Oct 1975.

[Rummel 76] W. D. Rummel, R. A. Rathke, P. J. Todd Jr., T. L. Tedrow, and S. J. Mullen, *Detection of Tightly Closed Flaws By Nondestructive Testing (NDT) Methods in Steel and Titanium*, NASA CR—151098, September 1976.

[Rummel 79] W. D. Rummel and F B Ross, *Reliability of NDI of Aircraft Engine Phase I,* MCR-79-678, Nov 1979.

[Rummel 81] W. Rummel, S. J. Mullen, F. B. Ross, and J. J. Jezek, *Reliability of NDI of Aircraft Engine Phase II*, SA-ALC/MM-772, Jun 1981.

[Rummel 82] W. D. Rummel, *Recommend Practice for Demonstration of NDE Reliability on Aircraft Production Parts*, Materials Evaluation, Vol. 40, pp. 922-932, 1982.

[Rummel 82a] W. D. Rummel, S. J. Mullen, B. K. Christner, F. B. Ross, and R. E. Muthart, *Assessment of TF-30 Fan Blade Inspection*, SA-ALC/MM-7881, October 1982.

[Rummel 82b] W. D. Rummel, S. J. Mullen, R. E. Muthart, F. B. Ross, and K. B. Christner, *Evaluation of the Capabilities and Limitations of Fluorescent Penetrant Materials in Air Force Applications, Phase III Report,* SA—ALC/MM—7884, May 1983. (Phase I Report SA-ALC/MM-7580, June, 1981 and Phase II Report, SA-ALC/MM-7883, December, 1982)**.**

[Rummel 83] W. D. Rummel, D. O. Thompson, and D. E. Chimenti, *Considerations for Quantitative NDE and NDE Reliability Improvement*, Review of Progress in Quantitative NDE, Vol. 2, Plenum Press, New York, 1983.

[Rummel 84] W. Rummel, S. J. Mullen, B. K. Christner, F. B. Ross, and R. E. Muthart, *Reliability of NDI of Aircraft Engine Phase IV,* SA-ALC/MM-8151, Jan 1984.

[Rummel 84a] W. Rummel, S. J. Mullen, F. B. Ross, and J. J. Jezek, *Reliability of NDI of Aircraft Engine Phase III,* SA-ALC/MM-7617, 1984.

[Rummel 86] W. D. Rummel, B. K. Christner, S. J. Mullen, and D. L. Long, *Characterization of Structural Assessment Testing,* SA-ALC/MMEI-1-86, January 1986.

[Rummel 86a] W. D. Rummel, B. K. Christner, S. J. Mullen, and D. L. Long, *Characterization of IBIS Fluorescent Penetrant Inspection Capabilities,* SA-ALC/MMEI-3-6, November 1986**.**

[Rummel 89] W. D. Rummel, G. L. Hardy, and T. D. Cooper, *Applications of NDE Reliability to Systems,* Metals Handbook, 9th edition, Vol. 17, pp. 674-688, 1989.

[Rummel 96] W. D. Rummel and G. A. Matzkanin, *NDE Capabilities Data-book,* NTIAC DB-95-02, May 1996.

[Rummel 97] W. D. Rummel, *Overview of NDE Capabilities in Aerospace Industry,* Proc. of European American Workshop on Determination of Reliability and Validation Methods on NDE, pp. 177-184, 1997.

[Rummel 97a] W. D. Rummel, *Overview of NDE Capabilities Assessments and POD Concepts,* Proceedings of 14th World Conference on NDT, Vol. 2, pp. 917-920, 1997.

[Rummel 99] W. D. Rummel, *NDE Reliability for Aging Aircraft,* Proc. of American European Workshop on NDI Reliability, Sept 1999.

[Rummel NY] W. D. Rummel, NASA MSFC-1249 - Currently being superseded by NASA STD - PO15.

[SAIC 98] Science Applications International Corporation, *C-141 Lower Inner Wing Spanwise Splice Joint Second Layer Fatigue Crack UT Inspection - Experimental Validation,* Final Report on WR-DEP Contract Engineering Task CET#TIEDM-96-1, May 1998.

[SAIC 99] Science Applications International Corporation, *Aging Aircraft Advanced Ultrasonic Inspection Techniques,* Final Report Vols. 1, 2 and 3, Report on WR-DEP Contract Engineering Task CET#TIEDM-96-1, Dec 1999.

[Sarkar 98] P. Sarkar, W. Q. Meeker, R. B. Thompson, T. A. Gray, and W. Junker, *POD Modeling for UT,* Review of Progress in Quantitative NDE, Vol. 17, pp. 2045-2052, 1998.

[Schroeder 88] J. E. Schroeder, D. W. Dunavant, and J. G. Godwin, *Recommendations for Improving AF NDI Technician Proficiency,* SwRI Project No. 17-7958-845, Dec 1988.

[Shepherd 95] W. T. Shepherd and C. G. Dury, *Human Factors in Aviation Maintenance: Current FAA Research,* DOT/FAA/AR-95/86, pp. 91-103, Jun 1995.

[Southworth 75] H. L. Southworth et al., *Practical Sensitivity Limits of Production NDT Methods in Aluminum and Steel,* AFML-TR-74-241, Mar 1975.

[Spanner 86] J. C. Spanner, *Human Reliability Impact on Inservice Inspection,* 8th Int. Conf. on NDE in Nuclear Industry, PNL-SA-13881, Oct 1986.

[Spencer 93] F. W. Spencer, G. Borgonovi, D. Roach, D. L. Schurman, and R. Smith, *Reliability Assessment at Airline Inspection Facilities,* DOT/FAA/CT-92/12, I, Mar 1993.

[Spencer 93a] F. W. Spencer, G. Borgonovi, D. Roach, D. L. Schurman, and R. Smith, *Reliability Assessment at Airline Inspection Facilities,* DOT/FAA/CT-92/12, II, May 1993.

[Spencer 94] F. W. Spencer and D. L. Schurman, *Reliability Assessment at Airline Inspection Facilities,* DOT/FAA/CT-92/12, III, Mar 1994.

[Spencer 96] F.W. Spencer, *Visual Inspection Reliability of Transport Aircraft,* NDE of Aging Aircraft, Airports and Aerospace Hardware, SPIE 2945, pp. 160-171, 1996.

[Spencer 96a] F.W. Spencer, *Visual Inspection Research Project,* Project Report, 1996.

[Spencer 97] F.W. Spencer, *Field NDI Reliability Study Designs to Incorporate Human Factors Issues,* Proc. of European American Workshop on Determination of Reliability and Validation Methods on NDE, pp. 273-282, 1997.

[Spencer 98] F.W. Spencer, *Fitting POD Curves to Single Inspector Hit/Miss data,* 1998.

[Sproat 72] W. H. Sproat, *Reliability Analysis of C-5A Pylon Inspections,* Lockheed Report LG72-ER-0106, Dec 1972.

[Sproat 79] W. H. Sproat and H Sharp, *Measurement of NDI Technician Proficiency on C-5 Wing Hotspot Inspections,* LG79ER0161, Aug 1979.

[Sproat 82] W. H. Sproat, *Air Force NDI Technician Proficiency Evaluation (ET, UT)-Final Report,* LG82ER0099, 1982.

[Sproat 84] W. H. Sproat and W. J. Rowe, *Ensuring Aircraft Structural Integrity Through NDE,* LG84WP7254-001, Sept 1984.

[Sproat 88] W. H. Sproat, J. M. Hamilton, and P. W. Hovey, *Engineering Services in Support of NDI Operations,* SA-ALC/MMEI/87-01, Oct 1988.

[Sproat 88a] W. H. Sproat, J. M. Hamilton, and P. W. Hovey, *Engineering Services in Support of NDI Operations,* LG88ER0078, Oct 1988.

[Sturges 86] D. J. Sturges, R. S. Gilmore, and P. W. Hovey, *Estimating POD for Sub Surface Ultrasonic Inspection,* Review of Progress in Quantitative NDE, Vol. 5A, pp. 929, 1986.

[Sturges 93] D. Sturges, *Approaches to Establishing Probability of Flaw Detection,* Advanced Composites Technology, Proc. of the 9th Annual ASM/ESD Advanced Composites Conference, pp. 668-679, Nov 1993.

[Summers 84] R. H. Summers, *NDI - Improved Capabilities of Technicians,* AFHRL-TP-83-63, 1984.

[Sweeting 95] T. Sweeting, *Statistical Models for NDE,* International Statistical Review, Vol. 63, No. 2, pp. 199-214, 1995.

[Swets 83] J. A. Swets, *Assessment of NDT Systems - Part I The relationship of True & False Detections; Part II Indices of Performance,* Materials Evaluation, Vol. 41, No. 11, pp. 1294-1303, 1983.

[Tanner 54] W. P. Tanner and J. A. Swets, *A Decision-making Theory of Visual Detection,* Psychology Review, Vol. 61, pp. 404-409, 1954.

[Taylor 89] T. T. Taylor, P. J. Heasler, and S. R. Doctor, *Use of ROC Curves in Measuring Inspection Performance,* Proc. of 12 WCNDT, Vol. 2, pp. 1059-1061, 1989.

[Taylor 99] T. Taylor and S. Doctor, *What is NDE Reliability Anyway?,* Second American-European Workshop on NDE Reliability, Boulder, CO 1999.

[Thompson 93] D. O. Thompson and L. W. Schmerr Jr, *Uses for Model Based POD Curves,* SPIE Vol. 2001, NDI of Aging Aircraft, pp. 121-132, 1993.

[Thompson 94] D. O. Thompson and L. W. Schmerr, *Role of Modeling in Determination of POD,* Advances in Signal Processing for NDE of Materials, Vol. 262, pp. 285-301, 1994.

[Thompson 97] R. B. Thompson and W. Q. Meeker, *A Methodology for POD Determination Incorporating Insight from Physical Models,* Proc of European American Workshop on Determination of Reliability and Validation Methods on NDE, pp. 107-114, 1997.

[Tober 99] G. Tober and W. B. Klemmt, *How is Reliability Used ? European Aerospace Viewpoint,* Second American-European Workshop on NDE Reliability, Boulder, CO, 1999.

[Triggs 86] T. J. Triggs, W. L. Rankin, R. W. Badalament, and J. C. Spanner, *Human Reliability Impact on Inservice Inspection,* NUREG/CR-4436, PNL 5741, BHARCC-400/85/016, Vol. 2, 1986.

[Wall 98] M. Wall, F. A. Wedgwood, and S. Burch, *Modeling of NDT Reliability (POD) and Applying Corrections for Human Factors,* ECNDT 98, Copenhagen, www.ndt.net/article/ecndt/relibil/325/325.htm, 1998.

[Walter 95] P. L. Walter and C. D. Smith, *The Aging Aircraft NDI Validation Center - A Resource for FAA and Industry,* DOT/FAA/AR-95/86, pp. 105-114, Jul 1995.

[Wheeler 86] W. A. Wheeler, W. L. Rankin, J. C. Spanner, R. W. Badalament, and T. T. Taylor, *Human Factors Study Conducted in Conjunction with a Mini Round Robin Assessment of UT Performance,* NUREG/CR-4600, PNL 5757, BHARCC-400/86/001, 1986.

[Yee 76] B.G.W. Yee et al., *Assessment of NDE Reliability DATA,* NASA CR 134991, Oct 1976.

INDEX OF MANUSCRIPTS