DOT/FAA/CT-92/12, I

FAA Technical Center
Atlantic City International Airport
N.J. 08405

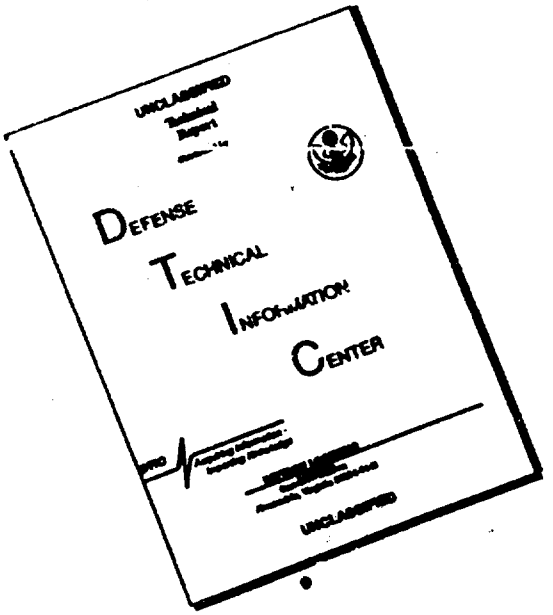
Reliability Assessment at Airline Inspection Facilities, Volume I: A Generic Protocol for Inspection Reliability Experiments

March 1993

Final Report

U.S. Department of Transportation
Federal Aviation Administration

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NOTICE

This document is disseminated under the sponsorship of the U. S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/CT-92/12, I		

4. Title and Subtitle

RELIABILITY ASSESSMENT AT AIRLINE INSPECTION FACILITIES VOL. I: A GENERIC PROTOCOL FOR INSPECTION RELIABILITY EXPERIMENTS

5. Report Date
March 1993

6. Performing Organization Code

8. Performing Organization Report No.

7. Author(s)
Floyd Spencer et al*

9. Performing Organization Name and Address

Sandia National Laboratories
Albuquerque, New Mexico

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

13. Type of Report and Period Covered
Final Report

12. Sponsoring Agency Name and Address
U.S. Department of Transportation
Federal Aviation Administration
Technical Center
Atlantic City International Airport, NJ 08405

14. Sponsoring Agency Code
ACD-220

15. Supplementary Notes

Chris Smith, FAA Technical Monitor
Pat Walter, SNL, Aging Aircraft Project Manager
Volume I of III

16. Abstract

*Giancarlo Borgonovi	SAIC
Dennis Roach	SNL
Don Schurman	SAIC
Ron Smith	AEA Technology

The Aging Aircraft NDI Development and Demonstration Center (AANC) at Sandia National Laboratories is charged by the FAA to support technology transfer, technology assessment, and technology validation. A key task facing the center is to establish a consistent and systematic methodology to assess the reliability of inspections through field experiments. This task is divided into three major areas: reliability of eddy current lap splice inspections at transport aircraft maintenance facilities, reliability of inspection at commuter aircraft maintenance facilities, and reliability of inspection associated with visual inspection of aircraft structural parts.

Volume I is the first document in a series of three describing the planning, execution, and results of an eddy current inspection field experiment. This document defines a generic protocol for inspection reliability experiments. It contains an introduction to the currently accepted forms of data analysis and presentation (Probability of Detection and Receiver Operating Characteristic curves) and a discussion of the factors that may affect inspection reliability.

17. Key Words	18. Distribution Statement
Nondestructive Inspection Eddy Current Inspection Inspection Reliability	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	37	

PREFACE

In August 1991, a major center with emphasis on validation of nondestructive inspection (NDI) techniques for aging aircraft was established at Sandia National Laboratories (SNL) by the Federal Aviation Administration (FAA). This center is called the Aging Aircraft NDI Development and Demonstration Center (AANC). The FAA Interagency Agreement, which established this center, provided the following summary tasking statement: "The task assignments will call for Sandia to support technology transfer, technology assessment, technology validation, data correlation, and automation adaptation as on-going processes." Key to accomplishing this tasking is the FAA/AANC Validation Center, which will reside in a hangar leased from the City of Albuquerque at the Albuquerque International Airport.

As one of its first projects AANC established a working group consisting of personnel from Sandia, Science Applications International Corporation (SAIC), and AEA Technology. The working group was charged with designing and implementing an experiment to quantify the reliability associated with airline eddy current inspections of lap splice joints.

The result of AANC's efforts is this three volume document, Reliability Assessment at Airline Inspection Facilities which details an experimental concept for inspection reliability assessment and a specific experiment designed to determine probability of detection (POD) curves associated with eddy current inspections. The experimental concept and eddy current experiment take into account human factor influences, which have not been fully addressed in past POD work. The result will be a better quantification of the reliability of the inspection techniques currently employed in the field. This will lead to better inputs for damage tolerance analysis and improved confidence in the specification of inspection intervals.

Because the FAA/AANC NDI Validation Center has been tasked to pursue other related NDI reliability experiments, the protocol for this experiment was developed first as a generic protocol then as a specific eddy current lap splice inspection protocol. The generic protocol is presented in Vol I: A Generic Protocol for Inspection Reliability Experiments, and the specific eddy current experiment protocol is presented in Vol II: Protocol for an Eddy Current Inspection Reliability Experiment. Because of the extent and duration of the experiment, the actual results of the experiment are presented separately in a third volume, Vol III: Results of an Eddy Current Inspection Reliability Experiment.

CONTENTS

Executive Summary	ix
1. Introduction	1
1.1 Purpose and Scope	1
1.2 Background	1
1.3 Reliability Assessment Elements	2
2. Experimental Plan	3
2.1 Functional Plan	3
2.2 Detailed Plan	3
2.3 Experimental Design	4
2.3.1 Test Variables	4
2.3.2 Flaw Distributions and Sizes	6
2.3.3 Number of Facilities and Inspections	7
2.4 Design and Manufacture of Experimental Specimens	8
2.4.1 Logical Design of Specimens	8
2.4.2 Flaw Characteristics	9
2.4.3 Experimental Setup	9
2.4.4 Specimen Identification	10
2.5 Specimen Characterization	10
2.6 Data Analysis	11
2.6.1 POD Analysis Using Logistic Regression	12
2.6.2 ROC Analysis Incorporating POFA	13
2.7 Protocols	15
2.7.1 Protocol Functions	15
2.7.2 Protocol Areas	16
2.8 Logistics	17
2.8.1 Assembly of Hardware and Dress Rehearsal	17
2.8.2 Safety Considerations	18
2.8.3 Scheduling	18
2.8.4 Storage and Shipment of Specimens	18
2.8.5 Field Adjustments to Loss or Alteration of Specimens	19
2.8.6 Post-experiment Archiving of Specimens	19
3. Implementation	20
3.1 Preparation	21
3.2 Experiment Execution	21
3.3 Data Qualification	21
4. Expected Results and Data Analysis	22
5. Preliminary Action Plan	24
References	25

LIST OF FIGURES

Figure 1	Reliability Assessment Elements	2
Figure 2	Example POD graph based on 3 inspections	13
Figure 3	Example ROC graph based on 5 confidence levels	15

LIST OF TABLES

Table 1	The number of facilities necessary	8
Table 2	Experiment Implementation Elements	21

NOMENCLATURE

a, flaw size	physical dimension of a flaw used in POD analysis models, usually a crack length.
AANC	Aging Aircraft NDI Development and Demonstration Center, run by Sandia National Laboratories for the Federal Aviation Administration.
ANOVA	Analysis of Variance, a statistical procedure for comparing the variability between groups to the variability within groups.
α, β	intercept and slope coefficients used in logistic regression equation. Other forms of the regresion equation may be parameterized by μ and σ, the mean and standard deviation.
baseline	a set of measurements performed under laboratory conditions on a given set of flaw specimens.
dress rehearsal	inspection tasks performed according to established protocols for the purpose of testing, in a realistic environment, all of the functions required to field a reliability assessment experiment.
ET	eddy current testing.
FAA	Federal Aviation Administration.
false alarm	an NDI response of having detected a flaw but at an inspection location where no flaw exists.
inspector	the person who applies an NDI technique, interprets the results, and decides whether a flaw is present.
MANOVA	Multivariate Analysis of Variance, a statistical procedure that generalizes the analysis of variance to consider two or more dependent variables simultaneously.
monitor	a person who observes and documents the results of inspections performed during an NDI reliability assessment experiment. This person will be familiar with the experimental goals and will have experience in human factors or the NDI technique under assessment.
NDI	nondestructive inspection, visual inspection is customarily excluded from being considered as NDI.
POD, POD(a)	probability of detection; as a function of flaw size, it is the fraction of flaws of nominal size, a, that are expected to be detected.

POFA	probability of false alarm
protocols	set of written procedures for conducting all activities required to implement a reliability assessment program.
PT	penetrant testing
QA	quality assurance
ROC	Receiver Operating Characteristic, a curve incorporating detection probabilities with probabilities of false alarms.
RT	radiographic testing.
SAIC	Science Applications International Corporation
SNL	Sandia National Laboratories
TSD	Theory of Signal Detection
UT	ultrasonic testing

EXECUTIVE SUMMARY

This document provides guidelines for planning and implementing field based experiments to determine the reliability of aircraft inspection methods.

The Aging Aircraft NDI Development and Demonstration Center (AANC) at Sandia National Laboratories is charged by the FAA to support technology transfer, technology assessment, and technology validation. A key task facing the center is to establish a consistent and systematic methodology to assess the reliability of inspections through field experiments. This task is divided into three major areas: Probability of detection (POD) in lap splice joints of transport aircraft, POD of cracks in commuter aircraft, and POD associated with visual inspection of aircraft structural parts. While planning the first of these activities, the AANC has developed the generic protocol for inspection reliability experiments that are described in this document.

These guidelines were derived from the experience of the AANC members, published work on previous reliability studies, and discussion with experts in the aviation industry. They are structured such that a detailed experimental plan for a particular inspection study can be developed quickly and thoroughly, yet maintain consistency with other studies. This structure will enable the AANC to build a consistent data base on the reliability of various inspection methods. From this data base, common trends can be determined and a basis established to make recommendations for improving inspection reliability.

The following planning, designing and implementing issues are described in this document:

Experimental planning . This covers the principal issues in the design of the reliability experiment including consideration of the test variables, flaw sizes and distributions, and number and type of facilities to be used. The essential features of the experimental specimens and their characterization are given. Types of data analysis for reliability studies are described. The protocols needed and the logistics for implementation are also presented.

Implementation of the experimental plan

Expected results and final data analysis

Plan of Action

The document also contains an introduction to the currently accepted forms of data analysis and presentation (POD and Receiver Operating Characteristic curves) and a discussion of the factors that may affect inspection reliability.

By conducting consistent, well planned and well executed reliability studies, the AANC and other organizations can make firm recommendations for improving inspection practices. Maintenance procedures can then be altered to improve performance at minimum cost and the inspection staff can gain a greater understanding of their jobs. This will result in a more informed and productive work force. This will lead to increased safety of operation, giving greater passenger confidence.

1. Introduction

1.1 Purpose and Scope

This document presents tasks and describes a systematic approach for assessing the reliability of nondestructive inspection (NDI) processes as they are employed in airplane maintenance and inspection facilities. The emphasis is on the tasks necessary for a viable assessment program and is not on specific NDI technologies. Thus, the approach described is applicable to a wide range of techniques including, but not limited to, eddy current testing (ET), ultrasonic testing (UT), fluorescent penetrant testing (PT), and radiographic testing (RT).

An NDI process encompasses the NDI hardware, procedures, inspectors, management, and the physical environment in which the inspection takes place. The reliability of such a process is defined in terms of two characteristics -- the probability of detection (POD) as a function of specified type of flaw, and the probability of false alarms (POFA). Both characteristics are ideally determined under the conditions existing at airplane maintenance and inspection facilities.

Including the POD and the POFA reflects concern for both safety and economics. The POD measures the ability of the system to find flaws (safety) but ignores the propensity to generate unnecessary followups and delays resulting from false calls (economics). Therefore, a POFA measure is needed.

The approach defined in this document is for inspections that produce binary ("detect" or "no-detect") data. That is, data are in the form of flaw detection or non detection. They are ***not*** in the form of a quantitative signal response.

1.2 Background

Definitions of terms used in NDI reliability assessments and generic programs for assessments have been reported in the literature [1-4]. Several extensive studies have provided valuable background information about NDI performance characteristics [5-7]. One recurring theme in these studies is that a major source of variation in POD curves is inspection-to-inspection differences.

Inspection-to-inspection variance potentially results from many factors. Physical conditions, such as accessibility, light, and noise, can influence inspection results. Factors specific to individual inspectors, such as training, recentness of experience, and alertness, can also influence the inspection results. All factors that are characteristics of the inspector or that influence the inspector's ability are referred to as "human factors." Because the focus here is to define a process to evaluate the reliability of existing NDI systems across many facilities, human factors play a major role in defining a reliability assessment program. The handling of these factors must be addressed in the experimental design.

1.3 Reliability Assessment Elements

Figure 1 is a flow chart showing various elements that are to be included in a reliability assessment of the nature discussed here. These elements have been grouped into four major phases. They are:

- Experimental Plan
- Experiment Implementation
- Expected Results
- Plan of Action (or Reaction)

Each major phase will be discussed individually in sections 2 through 5.

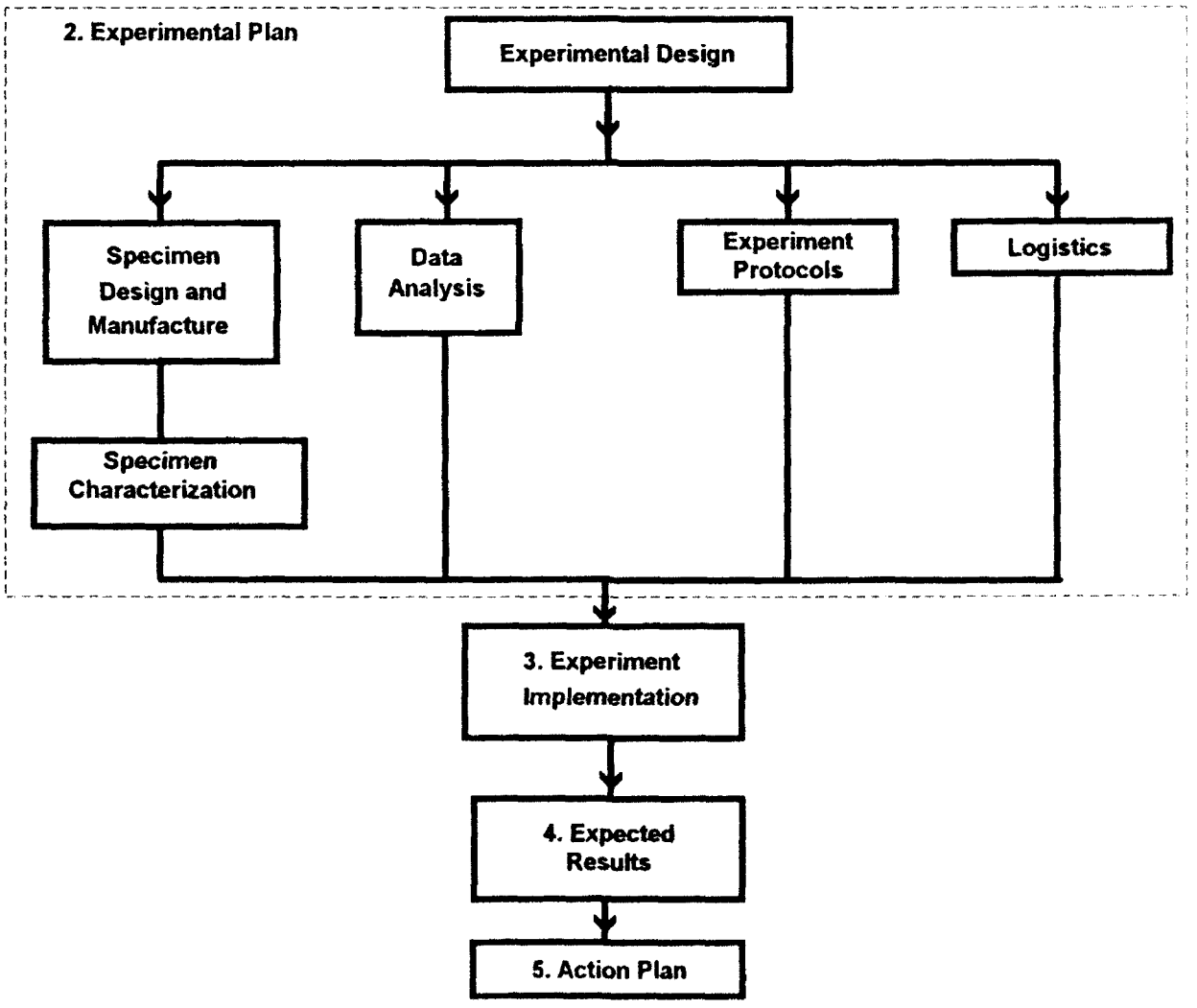


Figure 1. Reliability Assessment Elements

2. Experimental Plan

Prior to experimentation, an overall functional plan and a detailed plan for implementation must be developed.

2.1 Functional Plan

The functional plan spells out the missions and goals of the research, the methodology for creating the hardware to be used, the methods of observing or collecting the human factors of the test situation, the methods of implementing the research program in the field, and the structure and methods of data analysis.

The most important elements of any experimental plan for reliability assessment are the definition of the system being assessed and the statement of goals for the assessment. These are not only fundamental to the statistical design but they provide the only means by which the adequacy of the experimental plan can be judged. When combined with these essential elements, this document is meant to serve as the basis for a functional plan for NDI reliability assessment programs taking place at airplane maintenance and inspection facilities.

2.2 Detailed Plan

The detailed plan must list the steps that must be taken to fulfill the missions and goals of the functional plan. It should be a road map showing how to get from the overall functional plan to the results. The detailed plan must be specific enough that anyone familiar with the technology and the plan methodology will be able to perform the research exactly as the originators of the plan intended.

To finalize a detailed plan, the following activities must be performed:

- Determine experimental design.
- Agree upon designs and contracts for flaw specimen fabrication and characterization.
- Select data analysis procedures and methods.
- Outline materials to be used in briefing facility managers, training data collectors and (if necessary) inspectors, and collecting data.
- Develop schedules that include the order of steps, which can be done in parallel, and which must be done sequentially.

Activities of the detailed plan are discussed further in sections 2.3 through 2.7.

A **contingency plan** should be developed for unexpected situations. Controversy exists over the level of detail to be included in this backup plan. On the one hand, over-planning for contingency often leads field researchers to attempt force-fit solutions to problems that were not foreseen. On the other hand, failure to develop a contingency plan would mean that the first major problem might well collapse the whole program. The usual compromise is to predict as many contingencies as possible, based on prior experience and logical analysis, then plan for the most catastrophic or the most probable of these. There are typically only one or two backup plans. Given full communication among the principals involved and a clear chain

of command, unforeseen problems can usually be solved successfully. In these cases, the data collectors continue collecting data while the analysts develop new plans, that have a solid scientific basis, to deal with the changes in procedure or materials.

2.3 Experimental Design

After the NDI system for which a reliability assessment program is being defined has been specified and specific goals for the program stated, an experiment can be designed to meet those goals. General guidelines are presented for specifying test variables and flaw characteristics in the following subsections. It is recommended that a person with knowledge in the area of statistical experimental design be consulted in defining an experimental program.

Usually, a major goal of an NDI reliability assessment is to determine how effective an NDI system is under representative conditions. There will be variation in system response (as reflected by POD and POFA) induced by the variations in conditions at the time of application. The idea behind a designed experiment is to recognize the potential sources of variation and to explicitly specify how these are to be treated during the time that data are being taken.

Because reliability is usually characterized over a range of flaw sizes, the number of flaws and their sizes are elements of the experimental design.

2.3.1 Test Variables

To achieve a well planned experiment, it is necessary to identify those variables that potentially influence the results from the NDI system under study. All the variables identified may not ultimately be included in a test matrix, but by identifying them, one is in a better position to assess how "representative" the test conditions are to actual conditions.

In generating a list of potential test variables, the following general areas should be considered.

1. Facility Characteristics: This class of variables include the environmental characteristics that could influence the inspection results. Dim lighting in the inspection area could cause an inspector to make an inaccurate meter reading. General background noise could mask an audible alarm from an inspection device. A cold drafty environment could cause a rushed inspection and thereby influence the results. The environment could also influence the general alertness of an inspector.

2. Inspector: It has been observed in many applications that the human doing the inspection is the most significant source of variation in reliability. However, there may be identifiable factors with respect to the inspectors that could influence results. Specific candidate factors include training in the NDI techniques used and the recentness of experience. Other candidate factors might be age and physical characteristics of the inspector or the time required for the inspection task.

3. *Equipment:* The equipment used can cause reliability results to vary. In eddy current testing this category of variables would include the equipment type (for example, meter versus impedance plane), probe type, and the calibration piece used. In an NDI process such as pe: :trant testing, it would include the types of penetrants, emulsifiers, and developers, as well as the type of penetrant reader.

4. Inspection Process Variables: The inspection process may include variables that are not specifically controlled in procedures. These variables may be specified in ranges (such as eddy current inspections at 20 to 30 kHz). Inspection process variables (such as dwell times, scan rates, and frequencies) will be specific to the NDI system considered and may not be explicitly called out in procedures. Calibrations and how they are performed should be considered as a potential major source of variation in the inspection process.

5. Procedures: The procedures for a particular inspection may differ from facility to facility. Reliability changes that result from different procedures could be the result of different process variables within the procedures. Some written procedures may be so specific that they allow little variation regardless of who performs the inspection. Other procedures may be less specific and thereby allow more variation between inspections. Some of these procedural differences will be reflected in the inspection process variables. The structure of management and supervision within facilities could also be considered part of the procedural differences that impact inspections.

After the variables expected to have an influence on the reliability of an NDI system are determined, a decision can be made as to how to integrate them into the overall experimental design. The variables will fall into one of two general categories with respect to the experimental design: controlled or uncontrolled.

A controlled variable is one in which the values for that variable are specified with each inspection. For example, if the experimental test plan specified that half of the eddy current inspections were to be done with probe type "A" and half were to be done with probe type "B", then probe type would be considered the controlled variable.

An uncontrolled variable, on the other hand, is one in which the value is not specified in the experimental plan; rather, it takes on the condition or value at the individual facility. Thus, although the background noise level at a facility may influence inspection results, if no attempt is made to set the levels of noise during inspections, then this variable would be considered as an uncontrolled variable with respect to the experimental design.

Even though a variable is uncontrolled with respect to the experimental design, its value at the time of any one inspection can be recorded. Having the values associated with uncontrolled variables may prove helpful in the analysis of the data. This is addressed in more detail in Section 2.6 - Data Analysis.

The reason for controlling the levels of certain of the variables is related to the goals of the assessment program. The major goal 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.

For example, the primary goal of a program might be to assess the reliability of high frequency eddy currents in detecting surface cracks emanating from under rivets at specified inspection sites on a specified aircraft. However, it is recognized that different eddy current equipment is used in carrying out the inspections and that there are several procedures that can be followed. Therefore, a secondary goal might be to quantify the effects that the use of different types of equipment and the use of different procedures have on the overall reliability. Thus, one has to address how these two factors (equipment types and procedures) are to be incorporated into the experimental plan.

In designing an experiment where several factors are to be controlled, care must be taken to assign the values of each factor with each inspection in such a way as to assure that the factor effects can be estimated. In the probe type example above, assume the experiment is performed using two types of equipment, A and B, and two procedures, 1 and 2. If equipment type A was always used with procedure 1 and equipment type B was always used with procedure 2 then it would not be known if observed differences were due to the equipment or the procedures. On the other hand, if the trials were set up so that inspections occurred using each of the equipment types combined with each of the procedures, then not only could the effect of each factor (equipment type and procedures) be estimated, but one could test whether the effect was influenced by the presence of the other factor (that is, whether an interaction exists).

The above example is simple in that only two variables, each with two values, were considered. It is likely that many more than two variables may be of interest, some with more than two values. In such a case, a similar approach is required, but implementation becomes more complex; therefore, it is strongly suggested that a person with a background in experimental design be consulted.

2.3.2 Flaw Distributions and Sizes

Reliability assessments for NDI are generally in the form of POD(a) curves, which plot POD as a function of "a", where "a" is flaw size. The precision of any estimate of the POD(a) curve depends in part on the number of flaws included in the assessment program as well as the distribution of flaw sizes. The following discussion and recommendations are based on general guidelines that have been presented in the literature [1,8].

POD(a) curves in the form of models of linear log odds (See section 2.6. Data Analysis) have been shown to adequately fit NDI reliability data [8]. In this model the natural logarithm of {POD(a)/[1-POD(a)]} is expressed as a linear function of the natural logarithm of "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 that region where the POD curve is increasing.

A given set of specimens may be used in various NDI experiments, each experiment involving different types of NDI equipment. The region of flaw sizes providing the best information for each experiment is likely to differ. To accommodate multiple uses of the 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 [8]. Doing so would result in a distribution that favors the smaller sizes over the larger sizes.

A single POD curve as a function of flaw size represents an average of the detection probabilities for cracks of size "a". That is, it is not expected that two flaws of the same size will necessarily have the same detection probability associated with them. Therefore, the flaws included in an assessment that cover a range of flaw sizes also provide for an estimate of how variable POD(a) is for a fixed "a". Experience has shown that 30 flaws, with flaw sizes covering the region of interest (10th percentile to 90th percentile), are usually sufficient to obtain reasonably precise estimates of the POD curve. 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 [1,3,8].

Note that if flaw characteristics other than size are believed to have a substantial impact on the reliability assessment, then the above guidelines could be applied to each type of flaw being included in the assessment program.

It is also of interest to determine the propensity of an NDI system (including inspector) to make false calls. Estimates of POFA are based on the number of times an unflawed inspection site is called out as having a flaw. Therefore, any inspection task must contain unflawed inspection sites on which an estimate of POFA can be based. The unflawed inspection sites should be interspersed with the flawed sites so that the inspector has no reason to believe a site is flawed or unflawed before the inspection takes place. If possible, it would be desirable to make the ratio of flaws to unflawed sites mimic that ratio that inspectors would normally see. In most field applications, however, inspectors are likely to see very few flaws. An experiment with very few flaws may be unable to provide sufficient statistical power to get meaningful results. If the ratio of flawed to unflawed inspection sites cannot mimic "reality" because of too few flaws or excessive inspection time per inspector, it is recommended that the experiment contain about 3 times as many unflawed sites as flawed sites [3,8]. 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.

2.3.3 Number of Facilities and Inspections

The number of facilities and the total number of inspections obtained are also part of the experimental design.

It was pointed out earlier that many factors specific to the facilities may influence the reliability results. This includes procedural and training differences that might exist between facilities. For estimating facility-to-facility differences, each facility represents a single data point regardless of the number of inspections taking place within a facility. For this reason, the adequacy of a given number of facilities has to be considered on the basis of the total population of facilities and the specific goals of the reliability assessment program.

One way to gage the number of facilities that should be included in the experiment is to determine the number that would have to be sampled in order to achieve a given probability of obtaining at least one extreme facility. For example, Table 1 contains the number of facilities that should be included in order to have a 90% (or 95%) probability that at least one of the most extreme 10% of the facilities is included in the sample. From Table 2-1 it is seen that if there were a total of 30 facilities, then by randomly choosing 16, one is 90% confident that the sample includes at least one of the 3 (10% of 30) most extreme facilities.

The sample sizes of Table 2-1 are based upon random sampling. The criterion is to have a reasonable chance of including the range of variation in the sample. If information exists about likely causes of variation, it may be possible to choose "judgment" samples to reflect the variation, and thereby cut back on the number of required facilities.

Reference 3 recommends that at least five inspections be done at each facility. However, the number of inspections required for each facility may be driven by the experimental design with respect to the controlled test variables. If the experimental design contains eight different layouts, then it would be desirable to obtain eight different inspections at each facility so that no one inspector would have to inspect more than once. An inspector may perform more than one inspection, but the times of the inspections should be separated sufficiently to minimize the chances that the inspector has a "memory" associated with the first inspection.

Table 1. The number of facilities necessary to include at least one of the most extreme 10% at specified confidence level.

Total # facilities	Confidence level	
	90%	95%
10	9	10
20	14	15
30	16	18
40	17	20
50	18	21

One way to relate POD to the total numbers of inspections is as follows. The flaw size that has a 0.9 probability of detection is often estimated. Note that if, independent of inspector, a flaw has a probability of detection of 0.9 and a total of 30 inspections were performed, the probability of an estimate based on the results of the 30 inspections being as high as 1.0 or as low as 0.8 is about 0.11. More inspections would decrease the chances of these extreme estimates even more. (Or tighten the interval around 0.9 for the likely estimates.) Thus, it is recommended that the total number of inspections across facilities be at least 30.

2.4 Design and Manufacture of Experimental Specimens

When in situ experiments are carried out, the test specimens must accurately simulate all of the aspects that are critical to obtaining valid data. The specimens must not only model the structural points of interest, but they must also represent the global geometry as it normally presents itself to the inspector. Naturally, design tradeoffs will be present; however, these tradeoffs can be assessed only after cost, time, and prioritized experimental goals have been clearly established. With this procedure, the study can be carried out in a manner that does not compromise the experimental results. A number of factors play a part in achieving a realistic simulation of in-field aircraft inspections. These are listed and addressed in the following sections.

2.4.1 Logical Design of Specimens

Structural Detail and Layout. Although it is possible to recreate or use actual portions of the aircraft structure in its entirety, the purpose of the particular experiment may eliminate the need for some of the structure. For example, if a visual inspection of the outer skin is being assessed, there is no need to produce a structure that contains tear straps, stringers, or other subsurface structural components. However, it would be prudent to simulate reality as much as is possible in that portion of the structure that is directly involved in the inspection activity. Some experienced inspectors may, for instance, take cues from seemingly insignificant structural details.

The specimens should be sized so that they include a representative portion of the area to be inspected. Obviously, it would be desirable to build up an entire fuselage or wing or tail assembly; however, shipping

and assembly logistics, which are compounded by multi-site experiments, warrant that the specimen dimensions be minimized. These two conflicting requirements mean that design tradeoffs must be made. Human factors considerations and issues surrounding the variables being measured play a major role in assessing these tradeoffs and determining when the experiment has been compromised.

Specimen Manufacture. This is an important item since it normally has the greatest effect on the structural detail and layout of the specimens. Limitations established here will determine what type of specimen design is feasible. Cost and time are primary considerations in this phase of the specimen design. If, for example, specialized machinery must be designed in order to produce the specimens, then a significant investment of time and money may be required. Conversely, if existing equipment can be used to produce the specimens with the necessary characteristics (for example, size, shape, strength, flaw type, and flaw density), then this specimen design is preferable. It is important to note that numerous unknowns must be dealt with whenever customized machinery or new methods are used to produce test specimens. The end result may be specimens that are unacceptably different from the design plan.

2.4.2 Flaw Characteristics

In assessing any nondestructive evaluation technique or procedure, it is necessary that the specimens contain a number of known flaws. Two primary issues must be considered when deciding how to produce flaws in test specimens. These are flaw density and "real" versus artificial flaws.

Flaw Density. One method of producing flaws in test specimens is to control the number and distribution of flaw sizes in order to arrive at a statistically desirable experiment (as discussed in Section 2.3). Another way flaws can be generated is by using load levels and conditions that mimic those encountered on an actual aircraft. The end result will be a "natural" distribution of flaws. Any given experiment may include both "natural" and a statistically desirable flaw distribution.

"Real" versus Artificial Flaws. The issue to be addressed here is whether to produce the flaws using cyclic loading methods (that is fatigue the test specimens) or to actually machine artificial flaws into the specimens (for example, notches). Machined flaws have different characteristics than naturally occurring flaws, but this method allows for very precise placement, number, size, and orientation of flaws. Thus, it can be seen that a great deal of control is achieved; however, the experiment planners must determine if either the inspection method or the test results would be adversely affected by the use of artificial, machined flaws.

2.4.3 Experimental Setup

The overall presentation of the experiment is one of the most important factors in achieving inspector "buy-in." That is, the participant should feel that the experiment is quite similar to the actual inspection being modeled. Human factors considerations are critical to this area because inspector's data can vary greatly if the experiment is not presented as realistically as possible. Obviously, experimental limitations do not allow for an exact duplication of an airplane; however, every effort must be made to minimize the effect of these limitations on the experiment.

Use of an Airplane as a Test Specimen. Depending on the goals of the experiment, it may be possible to use an actual airplane, or a portion thereof, to conduct the tests. In experiments where numerous specimens are combined to model an inspection area of an airplane, the use of real hardware is normally not desirable. This is due to the following reasons: (1) it is very difficult to obtain the flaw density that is required to obtain meaningful results, (2) it is difficult to determine the flaw characteristics of the structure

after it has been assembled, especially when considering the number of structural members in a real airplane, (3) the added complexity of a complete aircraft structure is often not necessary, and (4) what is most important, without carefully produced test specimens, much of the experimental control is taken out of the hands of the planners.

Support Structure for Specimens. If an actual airplane is not used, then it may be necessary to mount the test specimens on some type of support structure or frame. The frame must adequately model the geometry of the chosen inspection area and must contain all of the features that are necessary to assure inspector "buy-in" and to faithfully reproduce the inspector/equipment/task interfaces. Key items to consider when designing a frame are (1) accessibility issues (for example, do not position the inspection area with easy external access if in reality the inspector has to perform his task from an awkward position) and (2) visual cues that will constantly remind the inspector that the experimental setup is not a real airplane.

The frame must also be designed with adequate consideration of cost, shipping, specimen mounting and alignment, and ease of assembly and disassembly. This is especially important if the experiment is to be carried out at a number of inspection facilities. One should also consider the possible re-use of a frame in follow-on experimentation, by incorporating some latitude in the set-up of the frame structure.

2.4.4 Specimen Identification

A methodology must be developed that provides a way to clearly identify and track each test specimen. This allows each specimen to be linked to its flaw characterization records. Without this link, the performance of the inspectors could not be assessed. The identification also allows the test specimens to be rearranged in an organized fashion so that different parameters can be studied during the multiple inspections.

The identification of a test specimen should be transparent to the inspectors. That is, if possible, the capability should exist to ask for a re-inspection by a given inspector without that inspector knowing he is inspecting a specimen that he has previously inspected.

2.5 Specimen Characterization

The purpose of specimen characterization is to confirm the size of the defects and subsequently enter the information in the data base. Estimating POD(a) requires that the flaws in the test specimens be accurately measured and characterized. Even though the technique being evaluated may be capable of providing a characterization of the specimens, other methods should be considered (including destructive methods on samples).

The characterization phase can be done in conjunction with the manufacture of the specimens. This approach has the benefit of providing an opportunity for feedback to be incorporated in the fabrication process.

Because the initial characterization is so important, independent determination of the dimensional properties of the flaws by several methods should be considered. For example, cracks can be measured by several methods. Visual microscopic measurements can be used. Laser profiling can be automated and can produce images that can be processed and displayed. Liquid penetrants are also standard in detecting surface cracks. In cases where direct visual observation may not be possible, ultrasonic imaging may prove to be valuable.

To increase the confidence in the previous characterizations by NDI methods, several specimens should be cross-sectioned. The surface at the cross section can then be photographed and the actual dimensions of the crack determined. Because making the samples is costly and because they have potential value for additional measurements using other techniques, the number of specimens examined destructively should be kept to a minimum.

It should be clear from the above discussion that characterization of the samples in connection with a POD experiment will not be a one-time operation, but rather a continuous activity taking place before, during, and after the experiment. To properly carry out the characterization one will have to:

- Provide a "home" for the specimens, that is, a place where they can be safeguarded and inspected as the need arises.
- Maintain a capability for nondestructive inspection which can be used as needed; for example, to verify that the specimens have not undergone changes.
- Provide adequate records on the results as well as on the procedures and the instrumentation used.

2.6 Data Analysis

The planned data analysis is an important and integral part of the experimental plan phase. Without a clear plan as to how the results of the experiment are to be analyzed, one cannot be effective in defining how data should be collected. The planned data analysis is therefore a necessary ingredient to accomplish many of the other aspects of a reliability assessment program. Most notably, the planned data analysis will affect the writing of protocols (section 2.7) and the writing of the overall functional plan (section 2.1).

The reliability of any NDI technique is often characterized by a probability of detection (POD) curve. This curve gives the probability of detection as a function of flaw size ("size" can be length, depth or other physical characteristics). Empirically, the probability associated with a flaw of a particular size can be estimated by the number of flaws detected (detects) divided by the number of inspection opportunities. It is recognized, however, that there can be many factors that influence the probability of detecting any one given flaw. It is the treatment of these other factors that has to be integrated into the planned data analysis. One accepted method is to use logistic regression, briefly discussed in section 2.6.1

There are problems associated with characterizing NDI reliability solely through POD curves. An arbitrarily high POD can be attained by simply saying that every possibility is a flaw. That is, a very high percentage of flaws can be correctly identified by simply setting a very low criterion for saying, "Yes, there is a flaw." This "bias toward yes" cannot be evaluated by calculating only the proportion of detects when a flawed site is being inspected, but must take into account the proportion of time a "detect" occurs when an unflawed site is inspected.

Stated another way, the reliability of any NDI technology should not be characterized solely by its ability to detect flaws. Therefore, a natural approach is to include some assessment of the probability that the NDI process will not give a false indication of a flaw. This leads to using the probability of false alarms (POFA) in addition to POD curves to characterize the reliability of an NDI technique.

The POD and POFA measures are integrated through a Receiver Operating Characteristic (ROC) curve (sometimes referred to as a Relative Operating Characteristic curve). The ROC curve plots the proportion

of correct "yes" responses against the proportion of incorrect "yes" responses. The ROC analysis is discussed in section 2.6.2.

In general, the data analysis and the discussions surrounding the data analysis need to address many points. The relationship of the experimental set-up to actual inspection conditions, the results of aggregating data, the relationship between POD and POFA estimates, and the identification and characterization of explanatory variables are examples of concerns to be addressed. The data analysis methods presented here are not inclusive by any means.

2.6.1 POD Analysis Using Logistic Regression

The log odds, or logistic regression, model has been used to analyze binary (detect or no-detect) data quite successfully in previous NDI reliability programs. See Berens [8]. The model equation is given by

$$\ln [POD(a)/(1 - POD(a))] = \alpha + \beta \cdot \ln(a) ,$$

where "a" is the flaw size.

The parameters α and β can be estimated by maximum likelihood methods. This can be done for a single inspector's data covering a range of flaws even though each flaw is either detected or not detected. That is, for any given flaw the best estimate for POD(a) is either 0 or 1, both of which make the left side of the above equation indeterminate. However, by having a range of flaws one can determine the α and β that maximize the likelihood of the particular sequence of 0's (non-detects) and 1's (detects) that was observed. Mathematical details are given in References 1 and 8.

How does the designed experiment discussed in Section 2.3 fit into the above model? One way of viewing the problem is as follows. For each inspector, maximum likelihood estimates for α and β are obtained, say α_e and β_e. The α_e's and β_e's can then be grouped and compared according to factors included in the original experiment. For example, consider that there were two types of equipment and two different procedures to follow. Then the set of α_e's and β_e's for the inspectors using equipment type 1 would be compared to the α_e's and β_e's for the inspectors using equipment type 2. Differences would be subjected to statistical tests for significance.

The comparisons discussed above can formally be built into the model statement by writing the right-hand side of the above equation to include parameters for the controlled test variables. For example, letting i= 1,2 for each of two types of equipment and j=1,2 for each of two procedures, the model becomes

$$\ln\{POD_{ij}(a)/[1 - POD_{ij}(a)]\} = \alpha_i + \delta_j + \beta \cdot \ln(a).$$

In this formulation, the effect of the type of equipment is reflected in the parameters α_i, and the effect of procedures is reflected through δ_j. Formal statistical tests are applied to determine if α_1 is significantly different from α_2, and similarly for the δ's. The above model can be expanded to include additional terms, depending both on the equipment and the procedure used. If such a term should significantly improve the model fit to the data, then this would be an indication of an interaction between the type of equipment and the procedure used.

The model formulation discussed above can be expanded to include parameters representing the effect of all the controlled test variables that were included in the original design. Parameters can also be included for factors not controlled in the original design but for which data were gathered at the time of inspection. For

example, the right-hand side of the log odds equation might be written as $\alpha + \beta \cdot \ln(a) + \gamma \cdot f$, where f is a quantitative measure of the light level existing at the time of inspection. Because f was not controlled, there is no guarantee that it will vary enough to give good estimates of the parameter γ.

An effective method of presenting the results of the POD analysis is to graph the proportion of inspections that result in detects against the flaw size and then overlay the POD function that results from the maximum likelihood estimates. An example is shown in Figure 2. This graphical technique can be applied to the individual inspections, but in that case each flaw would be graphed at either 0 or 1, according to whether it was detected.

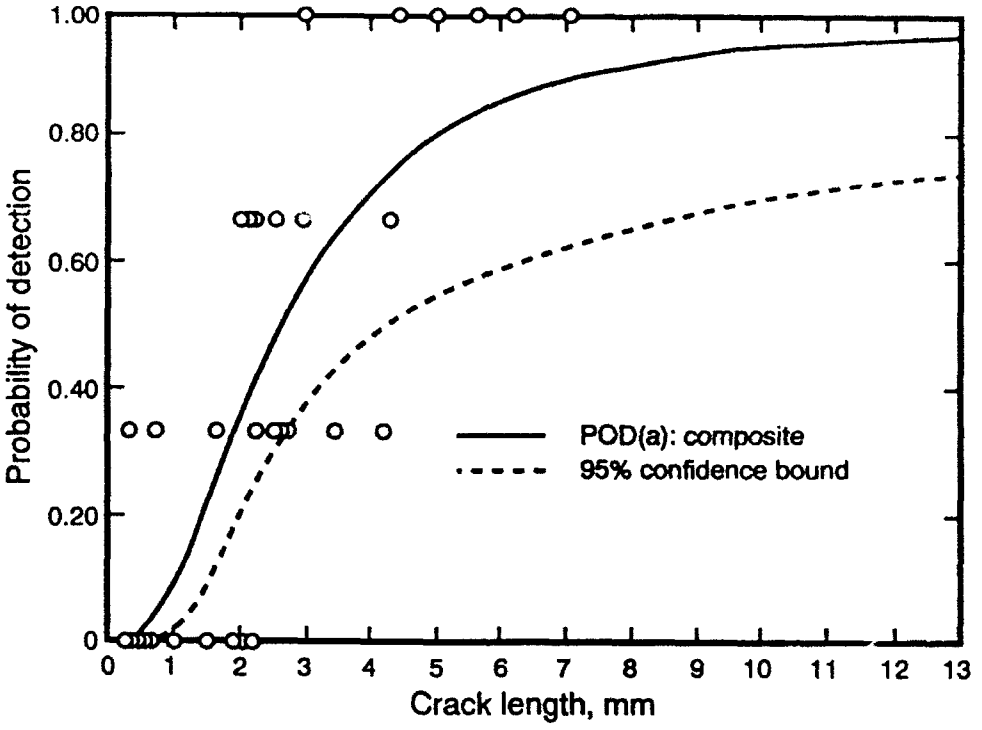


Figure 2. Example POD graph based on 3 inspections.

2.6.2 ROC Analysis Incorporating POFA

In the traditional POD analysis no consideration is given to the likelihood of a false alarm. One could modify the detect model to reflect a false alarm probability by considering the probability of an indication. Assume that POFA is a constant that applies for every inspection opportunity. The probability that an indication occurs is then given by

$$\text{Prob. of an indication} = \text{POFA} + (1 - \text{POFA}) \cdot \text{POD}(a).$$

This and other models are discussed in Swets [9]. The above model is basically a linear correction to POD, but Swets points out that prior studies of decision-making [10,11] have shown that humans do not simply hit or miss the target. For example, someone may incorrectly identify a scratch as a crack and be wrong,

but have been very unsure about identifying it as a crack. Are they as wrong as someone who answers, "Yes," without even looking? Are they as wrong as someone who looks and clearly mistakes the scratch for a crack and is very certain of their wrong answer? These questions pose a general concern about genuine "honest mistakes" and simple carelessness or incompetence.

These questions, and the subsequent investigation of these questions, led to the development of the Theory of Signal Detectability (TSD). The research into the general question of "How Wrong is Wrong?" led to the conclusion that simple linear corrections for guessing do not capture nearly enough of the information available in a human being's performance in detection tasks. TSD methods allow estimates of bias in answering, independent of correctness, and estimate the "true" ability of the person to detect/discriminate target situations (for example, cracks, splits).

TSD methods use "certainty" data to determine the criteria that the observers/inspectors are using to determine a "yes" response. Certainty can be determined by directly asking observers/inspectors what confidence rating they would give their answer, or it can be inferred through reaction-time data. In this research reaction-time data separate into reasonably distinct distributions. However, experiments that compared reaction time with verbal confidence ratings found that verbal confidence ratings were almost as reliable as reaction time distributions and were much easier to collect.

TSD measures are developed by plotting the proportion of correct "yes" responses against the proportion of incorrect "yes" responses for each level of confidence (see Figure 3). This plot, called a ROC Curve serves as the base for two TSD measures, called d' (d-prime) and beta. D-prime provides a good estimate of the "true" ability of the observer/inspector to detect/discriminate the target, independent of biases to answer in one way or another. Beta estimates the direction and magnitude of the observer/inspector's bias. See Swets [9].

TSD measures supplement, but do not replace, standard POD calculations. The TSD methods develop ROC measures that help determine the propensity to agree or disagree that a crack is present, independent of the actual situation (crack presence). A basic technique that can be used is to ask inspectors to tell how certain they are of their identification of cracks. In this way, the criterion level(s) they are using to say, "Yes, there is a crack here" can be determined. ROC curves will allow the determination of adjustments to be made to the POD curves to determine the actual ability of the inspectors to correctly detect cracks, given the materials, accessibility, and instruments that they are working with in each experiment.

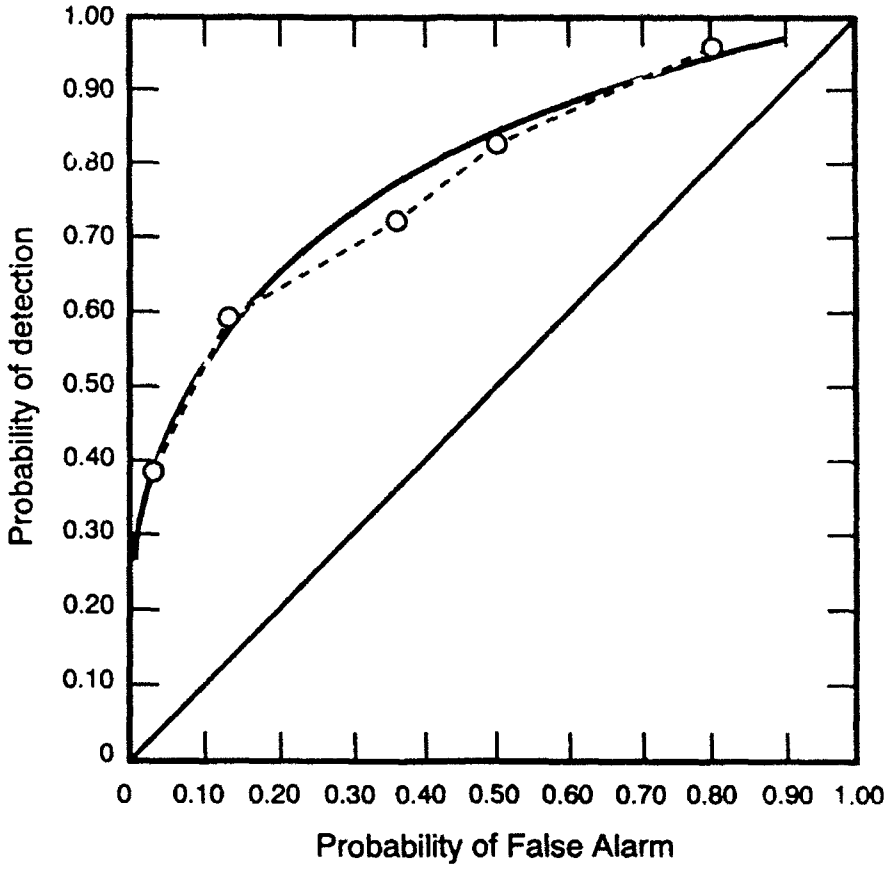


Figure 3. Example ROC graph based on 5 confidence levels

2.7 Protocols

It is important when carrying out complex experiments involving both experimental data gathering and human factors assessments that the experiment be regulated by a well defined set of protocols. The major areas for which protocols are necessary are discussed in the following paragraphs. The following paragraphs make reference to a monitor or monitors. These are those people who observe and document the results of inspections performed during an experiment. They will be familiar with the experimental goals and are likely to have experience in human factors or the NDI technique being assessed.

2.7.1 Protocol Functions

The primary functions of the protocols are to:

1. Assure that objectives of the experiment are implemented.

   It is vital that in major exercises (such as on-site POD trials) the purpose and objectives of the experiment be clearly specified at the outset and that they be in agreement with the aims of the sponsoring organization. This enables the test program and protocols to be defined in detail and accepted by the sponsor.

2. Assure that the experiment is carried out in a consistent manner.

The protocols ensure that the experimental program is adhered to as closely as possible by all concerned and that it conforms to the experimental plan.

3. Assure that recorded data are defined and gathered consistently.

The protocols ensure that the monitor knows what he is expected to do and what actions and information are required from him. They should ensure that the monitor has prepared thoroughly for individual test sequences, where different specimen arrangements are likely. They will specify the data to be recorded, its format and the recording medium and back-up requirements.

4. Assure that consistent information is given to the inspectors and their management.

Guidance is given to the monitor on the method and extent of the interactions with the inspector so that the monitor does not bias performance (for example, by passing on too much or too little information). Thus, the monitor effect is the same for each inspector. The inspector's procedures are essential and provide the basis by which the experimental results are assessed.

5. Assure that deviations from the experimental plan are dealt with effectively.

Guidance is required for the monitor to determine what course of action should be followed when program deviations appear likely (for instance, fast or slow inspections, interruptions, gross mistakes, or specimen damage). This will minimize the influence of unforeseen problems on the experimental results.

6. Assure that subsequent experiments can be carried out.

The protocols represent sound QA practices and provide the means by which the detail of the experimental program can be audited. This immediately implies that the experiments are reputable and hence any future data collected using the established protocols will be directly compatible and provide for objective comparisons.

2.7.2 Protocol Areas

The two primary sets of protocols are the activities and duties of the monitor, and the activities of the inspectors participating in the experiment.

The general areas for the monitor protocols are:

Operating test equipment
Briefings and questionnaires
Managing the tests
Test Specimen QA
Observing the inspector
Providing on-the-spot information
Controlling the documentation
Interacting with the inspector

Recording test conditions and environments
Recording inspection results
Analyzing the data

The general areas for the inspector protocols are:

Questionnaires
Structured briefings
Inspection procedures

2.8 Logistics

The logistical planning is the first stage of implementing an experiment. The demands of a complex experiment designed to be carried out at a number of facilities require a detailed plan of the logistics. This plan must ensure that the facilities are suitable for the experiment, that they can provide the resources necessary, and that the experiment can be safely and efficiently executed. There must also be a plan to schedule the experiment at various facilities in an orderly and cost-effective manner. A critical path analysis, with options for variance, should be developed.

There are several steps in the logistical plan. They can occur in both parallel and serial order. Typical steps are:

- assembly of hardware and dress rehearsal

- scheduling of experimental sessions

- safety considerations

- storage and shipment of experimental specimens

- field adjustments to loss or alteration of test specimens

- post-experiment archiving of test specimens

Each of the above steps in the logistical plan is discussed in the following sections.

2.8.1 Assembly of Hardware and Dress Rehearsal

Hardware assembly interacts with hardware design. However, the assembly procedures for the hardware must be practiced several times before the hardware is taken to a facility. This practice will occur during the "dress rehearsal" or trial run discussed below. The hardware should be designed for easy assembly and for ease of correct placement of the test specimens.

The amount of space required for assembly of the hardware and for use of the hardware must be ascertained beforehand and communicated to cooperating facilities to ensure that there will be adequate space to conduct the experiments. After the experiment arrives at the facility, environmental characteristics that could affect the inspector's performance (lighting, possible interferences, distractions, drafts, etc.) should be noted and logged. The impact of these characteristics on the inspection tasks should be

determined, and the physical layout should be adjusted to balance these characteristics with respect to other factors included in the experimental design.

Dress rehearsals of the test procedure must be held. The dress rehearsal should encompass the complete experiment, including setup, inspection, observation and data recording. Inspection staff should be familiar with the inspection task, and the monitors should be those designated to carry out the experiment in the field. The entire operation should be observed by trained specialists and a subsequent debriefing conducted with all those involved. This could lead to modifications of the procedures.

In these rehearsals, all monitors should become proficient in setup and takedown of the hardware. Special points of difficulty should be noted and, if possible, corrected.

2.8.2 Safety Considerations

Safety requirements and procedures that might impact placement, assembly, and use of the hardware should be determined at the time visits are scheduled (discussed below). Workplace restrictions must also be determined at the time visits are scheduled. Liability agreements must be addressed early.

2.8.3 Scheduling

It may be necessary to develop a tentative schedule at the same time that agreements to cooperate are being obtained from the facilities to be involved. A number of facilities that carry out the type of inspection under assessment should be contacted. At that time, briefings of the purpose and nature of the research at the facilities should be given to decision-level management. Also, requests should be made for facility representatives to provide information describing operations, particularly with respect to equipment, personnel, and operational organization.

Written confirmations of intent to participate should be obtained. Some facilities should be chosen as backup to others. Careful selection will allow minimum disruption due to unavailability of a facility at the scheduled time. As implied by this strategy, agreements should be obtained from more facilities (up to twice the number needed) than are actually to be visited.

The schedule should allow the monitors to return to the office for one week after every second or third week of monitoring. Since monitors may need to work two, or even three, shifts to obtain the required number of inspections, they will be working twelve to sixteen-hour days. Thus, they will need rest and recuperation. The off-time between facility visits can be profitably used in catching up with the paperwork and reducing and analyzing data. The early data analysis will also allow early detection of changes in the test specimens.

2.8.4 Storage and Shipment of Specimens

A custodian should be designated who has primary responsibility for the test specimens. The custodian should control, and be responsible for, the storage and shipment of the test specimens from the time of final characterization of the specimens through completion of experimental tasks. The custodian should arrange for secure storage facilities that will be unchanged for the duration of the experiment. The specimens should be logged into and out of those facilities, and access to the specimens should be limited in order to avoid mishap.

When weight and size restrictions permit, the specimens should be air-freighted when possible. Air freight assures both better schedule and physical integrity than overland freight. When there are breaks between facility visits (of more than two or three days), the specimens should be shipped back to central storage. Shipping arrangements should be the responsibility of the custodian at the storage facility, although the monitors are responsible for oversight of pickup and delivery on facility-to-facility shipments.

In the case of small specimens, the packaging should include a cataloging system for ease of inventory and subsequent specimen placement. Packaging should include appropriate mounting slots, padding, and such other protection as is necessary to preserve both the appearance and structural integrity of the specimens. If the specimens are to be archived for future use, the packaging should be designed such that permanent protection of specimens is provided.

2.8.5 Field Adjustments to Loss or Alteration of Specimens

The first line of defense against changes in the test specimens through vibration or rough handling is good packaging. For a case in which this defense does not work, the first alternative is to have backup specimens ready. The backup specimen should have, as close as is possible, the same number of flaws with the same characteristics as the specimen being replaced. The backups should be kept at the major storage location. Arrangements should be made for air shipment of the backup specimens as needed.

It will often be the case that the first intimation of a change will be a difference in the usual pattern of inspection responses. This difference will often not be detected until data is reduced at the end of a travel period. When a change is suspected, the suspect specimen should be sent back for characterization and a new one put in its place. Since the new specimen will not have exactly the characteristics of the old, the changed pattern must be logged and noted in the data files.

In the case of loss or destruction of test specimens, the adjustment will depend on the percentage losses relative to the total specimen group. Adjustment for any losses less than 50% can possibly be made by having inspectors go over the remaining specimens several times (in reverse order) to gain the required total number of responses. Some information will be lost with this procedure; however, adjustments in analysis can be made to salvage as much information as possible. This reduced data set may be preferable to terminating the whole experiment.

The contingency plan should address the above considerations. In addition, the above considerations depend on the experimental design. Therefore, a statistician or the original experimental designers should be consulted before plan alterations are made.

2.8.6 Post-experiment Archiving of Specimens

At the completion of an experiment, the test specimens should be characterized to verify that no alterations in flaw characteristics took place. The post-experiment characterization may be done on a sampling basis and include destructive testing. A preliminary survey of the experimental data can be used to guide the sampling process. In particular those samples where the inspection results are significantly different than would be expected based on the initial characterization might be targeted.

This characterization should be logged on the test specimen inventory records. Then the specimens should be stored, in their shipping containers, as a permanent library of well-characterized specimens.

3. Implementation

This section describes experiment implementation. The implementation process for an experiment is divided into three parts: preparation, experiment execution, and data qualification. The key elements of each are summarized in Table 2.

Table 2. Experiment Implementation Elements

Major Element	Key Factors
Preparation	Monitor Assignment Site Coordination Safety Specimen Acquisition Test Equipment Acquisition Shipping and Handling Experiment Integration Dress Rehearsal Storage
Experiment Execution	Site Introduction Site Preparation Inspector's Briefing Experiment Implementation Post Experiment Efforts
Data Qualification	Identification of all Data Sets Complete Data Sets Determination Data Integrity

3.1 Preparation

The preparation phase encompasses all of the implementation activities that need to be completed before the experiment is initiated. The specific efforts will be defined in the detailed experiment plan; however, as shown in Table 3.1, there are key factors that are common to all efforts of this type. In particular a dress rehearsal is strongly recommended because it provides a unique opportunity to try out, under reasonably realistic conditions, the equipment, procedures and protocol that will be employed in the experiments. The experimental plan can then be modified if problems in any of these areas are identified.

3.2 Experiment Execution

In this phase of implementation, the reliability assessment experiment is performed. The experiment execution efforts will be carried out in accordance with the requirements and procedures of the detailed experiment plan and the established protocol. The experiment execution may be impacted by unexpected situations, in which case a contingency plan should be in place (see Section 2.2 and 2.8.5). Since there are many sites and multiple repetitions of inspections, the execution activities, as summarized under Key Factors in Table 3-1, will include the efforts associated with these replications. Post-experiment efforts, therefore, also include the necessary disassembly, etc. in anticipation of movement to the next site.

3.3 Data Qualification

A vital part of a well performed experiment is the process of assuring the quality and completeness of the required data. As shown in Table 3-1, the focus of this effort is to ensure that the data acquired in the course of performing these experiments are valid, accurate, and properly identified.

4. Expected Results and Data Analysis

The need for planning the data analysis, as well as various accepted analysis methods was discussed in section 2.6. This section discusses in more detail the mechanisms by which the data are to be stored, made available, and presented. The need for reporting various types of analysis are discussed, but no mathematical details are given here. The mathematical basics are reported in References 1 and 8. The types of analysis needed can be performed by various commercially available software packages.

In a reliability assessment program that entails visiting many facilities, it is important that data analysis activities be done in parallel with the gathering of the data. Inspection results should be recorded on check sheets or other permanent media at the time the inspection occurs. That data can then be transferred to an electronic data base, the format of which will have been determined during the planning stage. The transfer to an electronic data base should be done in timely manner so that preliminary data analysis can take place. This preliminary data analysis can give indications of problems. For example, an unflawed site consistently being flagged as a flaw could indicate changes in the specimen.

If possible, the transcription of hard copy inspection results to an electronic data base should occur in the field. Both the electronic data base and the hard copy can then be sent (under separate cover) to the site where the analysis will be done. On-site analysis of inspection data by the personnel who will be monitoring the experiment is not recommended. The monitoring personnel should not divulge information concerning flaw distributions and characteristics, intentionally or unintentionally. To assure that flaw information is not given out, data files linking the specimens with specific flaw characteristics should not be available to the monitors during the implementation of the experiment.

By having multiple flaws that cover a wide range of sizes, POD curves can be fit to individual inspectors. These individual POD curves are basic to presenting the experimental results. By presenting POD curves for each inspector on the same plot, the total variation due to inspectors can be visually presented. Similarly, POD curves should be fit to the inspection results averaged across inspectors within a facility. A plot should be made with individual curves representing each facility. The result is a visual presentation of facility variation.

Plots showing inspector-to-inspector and facility-to-facility variation are fundamental. However, the significance of the observed variation depends on factors of the experimental design, including the number of flaws inspected and the number of inspections. The statistical significance should be assessed and reported.

Individual ROC curves can also be displayed on the same plot. Similarly, aggregated ROC curves by facility can also be put on a single plot. Thus, the same variation sources as were discussed for the POD analysis will be displayed in ROC presentations.

The results of ANOVA and MANOVA analyses on the α's and β's from individual POD fits and on various of the ROC measures should also be part of the basic reporting. The primary explanatory variables included in these analyses will be those that were controlled. However, exploratory analysis should be pursued using various of the observed conditions as possible explanatory factors.

An experiment that obtains data from many facilities, industry wide, will generate much interest. It is important to make the full data set of inspection results available for general distribution. This will result in independent analyses and will facilitate open discussion of the results. To this end, it is suggested that a

data base description document be prepared and that a flat PC-based ASCII file be made available for those requesting the data. This data base should contain no specific facility nor inspector information.

5. Preliminary Action Plan

The main efforts of any reliability assessment experiment are spent in planning, implementing, analyzing, determining POD estimates, and producing the final report. The process, however, is not complete until the initial goals of the program have been reviewed and compared with the program results. Normally, this is accomplished by making a list of recommendations, usually reaching certain conclusions, and recommending what elements require further investigation.

During the course of conducting a reliability assessment experiment, a great deal of information is collected and analyzed. Based on the data collected, it is not unlikely that certain courses of action can be recommended that could have an impact on the overall inspection process and improve its reliability. It is therefore strongly recommended that a Preliminary Action Plan be determined, based on the lessons learned and data collected while conducting the experiment.

This plan should consider, but not be limited to, actions pertaining to such inspection aspects as:

- Standardized calibration techniques for equipment, sensors, and calibration blocks
- Inspector training programs
- Standardized certification examinations
- Re-examinations
- Improvements in environmental conditions
- Improvements in staging
- Improvements in fixtures
- More meaningful data collection

Based on the information available, methods for implementing the above (and other identified) recommendations should be determined. A preliminary schedule of implementation and an order-of-magnitude cost analysis should also be prepared.

The benefits of action based on reliable inspection data are manifold. They should lead to increased safety of operation and give greater passenger confidence. By identifying the key reliability issues, maintenance procedures can be optimized to improve performance at minimum cost, and inspection staff can gain a greater understanding of their own jobs. The result will be a more informed and productive work force.

References

1. Annis, C., Berens, A., Bray, F., Erland, K., Hardy, G., Herron, W., and Hoppe, W., MIL-STD for USAF NDE System Reliability Assessment (Proposal), Draft No. 2, Aug. 1989.

2. Oelkers, E., and Holt, A.E., "An Approach to Program Planning for NDE Reliability," in Nondestructive Evaluation (NDE) Planning and Application, R.D. Streit, ed., ASME, 1989.

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

4. ASNT, "Recommended Practice for a Demonstration of Nondestructive Evaluation (NDE) Reliability on Aircraft Production Parts," Introduction by W.D. Rummel, Material Evaluation, 40, August 1982.

5. Lewis, W.H., Sproat, W.H., Dodd, B.D., and Hamilton, J.M., Reliability of Nondestructive Inspections: Final Report, SA-ACC/MME 76-6-38-1, San Antonio Logistics Air Command, 1978.

6. Rummel, W.D., Mullen, S.J., Christner, B.K., Ross, F.B., and Muthart, R.E., Reliability of Nondestructive Inspection of Aircraft Engine Components, SA-ALC/MM8151, San Antonio Logistics Air Command, 1984.

7. Bush, S.H., Reliability of Nondestructive Examination, NUREG/CR-3310, Vol. 1, U.S. Nuclear Regulatory Commission, Washington DC, 1983.

8. Berens, A.P., "NDE Reliability Data Analysis," in Metals Handbook, Volume 17, 9th Edition, ASM International, Materials Park, Ohio, 1988.

9. Swets, J.A., "Assessment of NDT Systems - Part I: The Relationship of True and False Detections" and "Assessment of NDT Systems - Part II: Indices of Performance" in Materials Evaluation, 41, October 1983.

10. Swets, J.A., and R.M. Pickett, Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, Academic Press, Inc. New York, NY, 1982.

11. Green, D.M., and J.A. Swets, Signal Detection Theory and Psychophysics, John Wiley, New York, NY, 1966; reprinted by Krieger Publishing Co., Melbourne, FL., 1974.

☆U.S. GOVERNMENT PRINTING OFFICE: 1993 - 704-078/80026