Introduction

The goal of this paper is to review the statistical methods used in the aerospace industries to evaluate NDE reliability. The techniques presented are consistent with the damage tolerant design and structural maintenance philosophies of the aerospace industry. The first part of this paper establishes the evaluation criteria and discusses the history of NDE reliability evaluations. The second part describes the state-of-the-art analysis methods through examples from the retirement for cause (RFC) inspection system evaluation. The last part of the paper discusses some techniques used to rate operator performance and deal with false calls.

Background

NDE and DTA

The Air Force requires the design of damage tolerant aircraft and manufacturers use tests and damage tolerance analyses (DTA) to satisfy this requirement and to make life predictions. The basic analytical approach to damage tolerant design involves fatigue crack growth predictions and is presented schematically in Figure 1. Predicted fatigue growth from an assumed initial flaw size, \( a_0 \), provides a bound on fatigue crack growth from all possible manufacturing defects taking into account quality control measures such as NDE. An inspection is required at half the predicted time to failure. After the in-service inspection the flaw size is reset to \( a_{NDE} \), a measure of field service NDE capability. The process is repeated throughout the life of the aircraft.

The reset flaw size is, conceptually, the largest flaw that can be missed at an inspection. Under the ideal inspection system, all flaws larger than \( a_{NDE} \) are detected while no flaws smaller than \( a_{NDE} \) are detected. Realistically, not all flaws of a given size are detected repeatedly by a given inspection system. The probability of detection (POD) is defined to be the proportion of times an inspection system will detect flaws of a given size. The idealized (POD) as a function of crack length is the step function illustrated in Figure 2. A POD curve which is more representative of realistic NDE capability is also shown in Figure 2.
The POD function serves as the basis for NDE detection capability evaluation in the aircraft industry. The POD is described as a function of crack size since life prediction calculations for damage tolerant structures are based on crack size. The role of NDE is to enhance life by finding cracks so the effectiveness of NDE should be evaluated by the sizes of flaws that can be found. The POD function provides a measure of how well flaws of all sizes can be found.

As with most safety critical applications, a conservative measure of NDE capability is desired. The exact capability cannot be known exactly because it can only be measured from limited test programs. The random sampling conducted in the test program along with the variability of the NDE process lead to errors in the estimated NDE capability. A confidence bound provides a conservative measure of NDE capability that accounts for
the random errors of the test program. The aircraft industry uses the 95
percent lower confidence of the POD function to estimate NDE capability.
The use of a lower confidence bound provides a quantifiably controlled risk
that the true POD function is lower than the indicated NDE capability.

History of Estimation Methods

The earliest methods of estimating the POD function involved the use
of Binomial statistics applied to find/no find data. Flaws were grouped by
size and the POD for the group was estimated as the number of finds divided
by the number of opportunities [1]. The confidence bound calculations with
this technique are inadequate because the confidence bound is highly
affected by the sample size. Various 'improvements' for smoothing the POD
function and confidence bound have been implemented through the years [2];
however these methods only provide approximate confidence bounds which are
overly conservative.

The statistical analysis performed as part of the 'Have Cracks Will
Travel' program [3] was the first instance in which an analytical function
was fit to estimate the POD function. In the 'Have Cracks' study each flaw
was inspected a large number of times allowing estimation of the detection
probability for each flaw. Several analytical functions were evaluated as
potential models for the detection probabilities.

Although better estimates of the POD curve were achieved, the Have
Cracks study still relied on Binomial statistics to calculate confidence
bounds. Berens and Hovey [4] re-analyzed the Have Cracks data and showed
that a linear log logistics analysis provided a more acceptable fit and the
correct confidence bound for the POD function. The linear log logistics
analysis is a popular biometric method commonly used to estimate the prob­
ability of success as a function of discrete and continuous variables (such
as dosage).

Current Methodology

The analyses conducted in [4] showed that the linear log logistics
function provided a good empirical fit to the NDE detection reliability
data from the Have Cracks program; however it did not provide a justifica­
the good fit provided by the logistic distribution function through a model
of the response signal of the NDE device as a function of crack length.

In many NDE systems, and automated systems in particular, a response
signal amplitude is used for flaw detection decisions. The response signal
is referred to as \( \hat{a} \) to distinguish it from the crack length, \( a \). If \( \hat{a} \)
is larger than the detection threshold the system gives a detection indication
while \( \hat{a} \) values smaller than the threshold are ignored. Figure 3 shows a
typical example of \( \hat{a} \) versus a data collected with the RFC eddy current
inspection system.

The data in Figure 3 were collected on titanium web bore specimens.
The horizontal lines represent the saturation limit of the system (top) and
the recording threshold for collecting the data (bottom). Data on the
horizontal lines are censored and the two mean trend lines represent two
methods of addressing the censored data. The dashed line was fit using
only the flaws that were found while the solid line provides the correct
fit to the data including the censored data. The shift in the mean trend
line due to ignoring the censored data can cause errors of 50 to 100 per­
cent in the estimated location of the POD function. The close correlation
between \( \hat{a} \) and \( a \) allows the system to make sharp discriminations of flaw
size for this specimen.
Fig. 3. The effect \( \hat{a} \) of censored values.

A method of estimating the POD function from \( \hat{a} \) versus \( a \) data was presented in [5]; which assumes a linear relationship between \( \ln(\hat{a}) \) and \( \ln(a) \) and that the scatter in \( \ln(\hat{a}) \) about the mean trend has a normal distribution. Figure 4 contains a schematic of how the percentiles of \( \hat{a} \) corresponding to the detection threshold are converted to POD values at each crack length. The locus of POD value - crack length pairs forms the POD curve.

The shape of the transformed POD function in Figure 4 is derived from the distribution of \( \hat{a} \) values. The response signal is assumed to have a log normal 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 and is preferred for the analysis of pass/fail data because it is analytically simpler. 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 analysis and the \( \hat{a} \) versus \( a \) analysis have become standard procedures for characterizing NDI reliability in the aircraft industry. They are used by the aircraft engine manufacturers and were also used in the recent evaluation of the NDE system developed in conjunction with the Air Force Retirement for Cause program. They have also recently been implemented by Battelle for the Air Force NDI Program Office at San Antonio Air Logistics Center for use in surveillance and control of Air Force laboratories and shops.

EXAMPLE RESULTS

The log odds analysis and the \( \hat{a} \) versus \( a \) analysis differ mainly in the type of data they were designed to handle. The log odds analysis was designed to evaluate data with binary (find or no find) responses from the NDE system. The \( \hat{a} \) versus \( a \) analysis can only be used when a response amplitude is used for detection and the values of the response amplitude from the test inspections are available. Since \( \hat{a} \) data can easily be reduced to binary responses, the pass/fail analysis can also be used when
the response amplitudes are available. The following paragraphs present examples of both the $\hat{a}$ versus $a$ and the pass/fail analyses applied to data collected from the RFC inspection system.

$\hat{a}$ Versus $a$ Analysis

One of the assumptions of the $\hat{a}$ versus $a$ analysis is that the scatter in $\ln \hat{a}$ about the linear fit has a normal distribution. Figure 5 illustrates a common test for normality applied to data in Figure 3. The residuals, $\ln \hat{a}$ minus the mean trend, are plotted on a normal probability scale. The high correlation between the normal scores and the residuals is an indication that the normality assumption is valid for this data. The $\hat{a}$ versus $a$ analysis has been tested on many data sets from automated NDI systems and the assumptions of the analysis method have been met adequately in each case. The log normal/log logistic POD model thus continues to be an acceptable model.

When the assumptions for the analysis have been verified, the estimates of the POD function and its confidence can be calculated. Plots of the estimated POD function and its confidence bound for the data of Figure 3 are contained in Figure 6. The results are typical of highly automated systems with a steep POD function and a tight confidence bound.
Pass/Fail Analysis

The $\hat{a}$ versus $a$ analysis is preferred to the pass/fail analysis; however the pass/fail analysis was also conducted on the RFC data to serve as a comparison. The POD function and confidence bound for the titanium web bore data calculated using the pass/fail analysis method are shown in Figure 7. A comparison of Figures 6 and 7 shows that the mean trends for the POD function are similar for the two analyses; however the confidence bound for the pass/fail analysis 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 two outcomes: pass or fail.

FALSE CALL EVALUATION

The POD function serves as the main evaluation tool for NDE capability in designing aircraft systems. The issue of false calls is not generally considered in the design phase because false calls do not directly affect the safety of the system. False calls are more of an economic or operational consideration and are discussed briefly.

Few techniques have been developed to evaluate false call rates because no clear or consensus definition of what constitutes a false call has been established. Calling an unflawed region flawed is of course a false call; however under some circumstances it could be advantageous not to detect small flaws that pose no threat to structural integrity. Depending on the definition of a false call different procedures would be required to estimate the false call rate. The one area of NDE evaluation in which false calls have been addressed is in operator performance evaluation.

Fig. 5. Test of normality for $\hat{a}$ residuals.
False calls are important in operator evaluations because an operator could find all flaws by calling all parts flawed whether or not the NDE system has given a positive indication. The two main tools for operator evaluation are the coefficient of contingency and the relative operating characteristic. Both of these tools measure the tradeoff between high POD and a low false call rate.

The coefficient of contingency is a measure of association between the presence or absence of a flaw and the operator/system indication. If the inspection results are independent of whether or not a flaw is present,
either the operator is basically guessing or the inspection system is inadequate. On the other hand if the operators decisions perfectly coincide with the true state, he and the system are a valuable asset to the Air Force. An example of the use of the coefficient of contingency is presented by Davis and Aguilar [6] in these proceedings.

The relative operating characteristic (ROC) is a plot of the detection percentage versus the false call percentage for a particular operator/system combination. The ROC plot facilitates establishing which inspectors are performing within specified bounds of detection and false call percentages. The ROC presents similar information to the coefficient of contingency in a visual format. Further information about the use of the ROC is given by Heasler [7] in these proceedings.

SUMMARY

The POD function is the main evaluation tool for NDE reliability in the aircraft industry. Two statistical techniques have been developed for estimating the POD function as a function of crack size. Linear log logistic analysis is used for pass/fail data while the $a$ versus $a$ analysis is more appropriate for data that includes the raw response signal for each flaw. The techniques available for estimating the false call rate are geared more toward operator performance evaluation than evaluation of the influence of NDE on the reliability of the system.

REFERENCES


