Probability of Rejection - In conformance with DNV OS F101

E.A. Ginzel [1], M. Matheson [2], B.Feher [3]
[1] Materials Research Institute, Waterloo, Ontario, Canada
[2] Eclipse Scientific Products, Waterloo, Ontario, Canada
[3] Research In Motion, Waterloo, Ontario, Canada

Abstract:
NDT use of the concept of "Probability of Detection" (POD) has been around since about the 1970s. Most recently the concept of "Probability of Rejection" has been used. A description of its premises and how it is easily adapted from the principles of POD are described here.

Keywords: Probability, Probability of Detection, Probability of Rejection, PoD, PoR

NDT Flaw Detection and Thresholding

By now most practical users of ultrasonic testing (UT) are aware that ultrasonic testing is method filled with variables. These can include but are not limited to;

- flaw size,
- flaw orientation,
- flaw surface texture,
- beam characteristics,
- position that the flaw is detected along the sound path,
- potential that the flaw was off-axis,
- anisotropic characteristics of the materials tested,
- coupling variation due to test surface texture,
- texture of the local surface conditions where a skip is required,
- weld cap and weld root geometry,
- mismatch conditions, etc.

Most UT inspection processes rely on amplitude of signals from flaws to initiate evaluation. Even TOFD requires that a flaw provide some indication of diffraction over the background "noise" level. Since most of the many variables encountered in UT result in deviation in registered amplitude from flaws, there can be a wide variability in what is evaluated in UT. Yet in spite of these variables, ultrasonic inspections of welds in pressure retaining components seem to have provided good quality products as evidenced by the relatively low numbers of failures.

However, it is likely that this is not because NDT codes are ideally suited to ensure that all critical flaws are detected and removed! In fact, when we consider some of the codes and the variables involved, it is perhaps more by good fortune than design that we have had success.

When using "acceptance criteria" in NDT, whatever the NDT operator sees is evaluated against the acceptance criteria. But this process is not as quantitative as engineers would seem

to have us believe. What is it that the NDT operator must "see"? The NDT operator must make a decision as to what is noise and what is a valid indication of a flaw. "Noise" is present in all NDT methods. Colour-contrast liquid penetrant inspection has speckled effects on a white background to contend with. Magnetic particle inspections have small field indications of particle clusters that result from small surface roughness variations. Radiography has the grain effects in the film emulsions or illusory effects of geometric contrasts. These methods tend to be far more qualitative than eddy current or ultrasonic methods when it comes to determining the threshold at which to evaluate an indication.

Ultrasonic testing and Eddy Current testing both use some form of electronic displays that can be configured to indicate a voltage displacement as the signal indicator. Both can be subject to noise as well. Eddy current can have small voltage variations that result from probe motion and ultrasonic testing can be plagued with scattered signals from coarse grains. But the operator in an eddy current or ultrasonic test generally has a requirement to maintain a signal to noise ratio based on some reference target. With a minimum signal-to-noise ratio the inspection can then be configured to indicate a clearly defined response from the reference target and a "threshold" is set above which the operator "evaluates" signals.

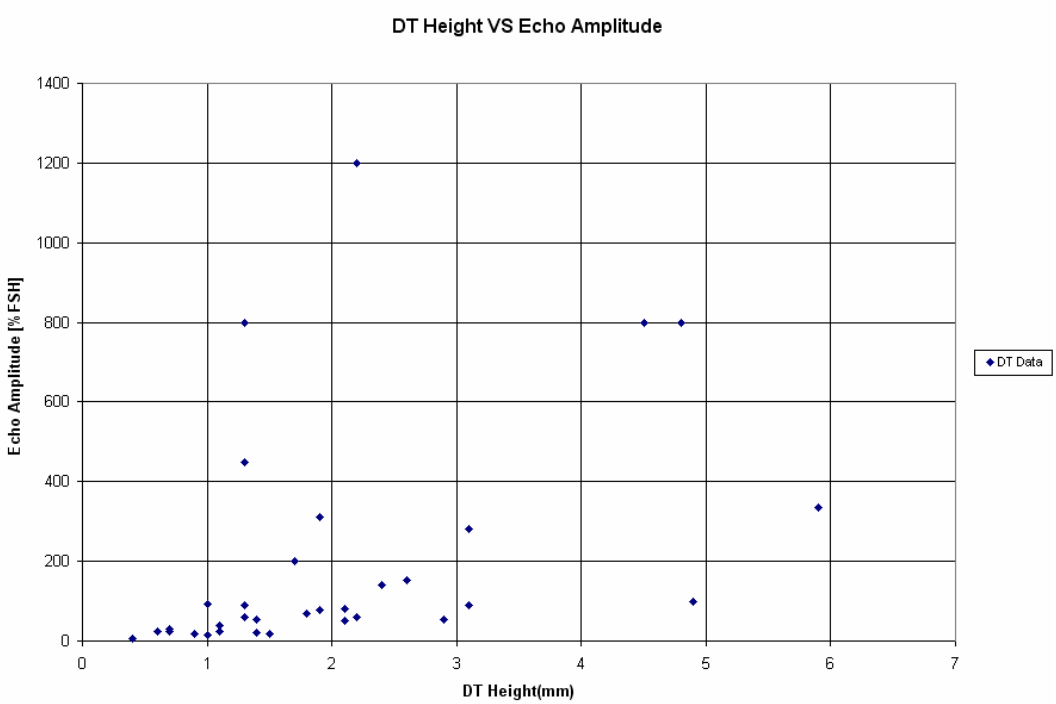
In ultrasonic testing this is typically done using a distance amplitude correction (DAC) curve. Signals that exceed some amplitude relative to the DAC are evaluated (typically for length but may also be assessed for vertical extent). Or in some cases, there is no dimensional assessment and the indication is simply considered unacceptable.

Underlying the evaluation process is the concept that all serious flaws will be identified as signals above the "threshold" and will not be confused with background noise.

This would seem to make the process a simple Boolean sorting; i.e. go/no-go. Anything above the threshold is bad and anything below is acceptable. For "workmanship enforcement" criteria this may be suitable. But workmanship criteria have no concern for the serviceability of the component or repercussions of the repair process. When the acceptance criteria are based on the mechanical properties of the materials used, the goal is invariably to discern if the flaw will be detrimental to service or not. This implies that there is some "critical flaw size" above which the inspection process must identify. Therefore all flaws that are greater than the "critical flaw size" must be identified (detected) and this is accomplished by producing a signal response above the critical flaw size "threshold".

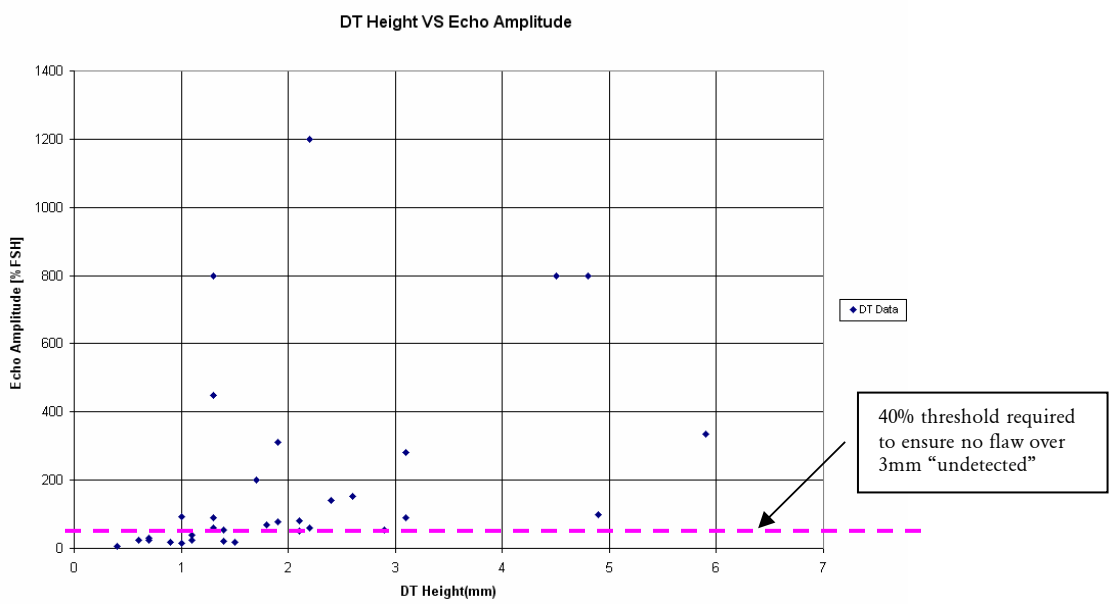
Instead of its initial purpose as a "prod" to make the welder pay close attention to their welding, the workmanship criteria used in many Codes have been warped into the idea that the threshold level is actually maintaining a structural function. Simplistic treatment of ultrasonic test-results equates repairing of flaws over a specified amplitude and length to structural integrity and safety. Figure 1 indicates a plot of amplitude versus flaw height (generally the critical flaw parameter in a fracture-mechanics analysis). This indicates the lack of relationship of flaw size to response (amplitude as a percentage of reference).

Figure 1 Flaw Height Versus Echo Amplitude



When this information is translated into fitness-for-purpose concepts and the critical flaw size is calculated for the application, a horizontal line can be drawn such that it ensures that all flaws over that size are above the line. This provides the threshold amplitude that ensures flaws over the critical flaw size are identified as “detected”. Only “detected” flaws are then sized and compared to the acceptance criteria.

Figure 2 Flaw Height Versus Echo Amplitude with Critical Flaw Size Threshold indicated



The link between amplitude and flaw size has long been the "ideal" in ultrasonic testing. The AVG system (DGS in English) has been made based on the ideal response of a perfect disk shaped reflector on the axis of an ultrasonic beam. But even its developers Krautkramer and Ermolov[1,2] caution against relying on such idealised conditions. Real flaws provide a scattering of echo amplitude responses. Graphs in Figures 1 and 2 illustrate that the response from any real flaw is therefore a probabilistic event.

Figure 2 also illustrates another problem. When the evaluation threshold or critical flaw size threshold approaches the noise level there is a risk that some critical flaws may be just under the threshold and some non-critical flaws will be unnecessarily identified as rejectable. When the threshold is set very low the transition between the acceptable and unacceptable condition is poorly defined and the rejection process becomes a random event. Random events provide a source for probability assessments.

Origins of POD and Pipeline Consruction Applications

Probability is a mathematical concept that is of great importance to NDT. With the many variables associated with amplitude fluctuation, the nature of "detection" as an event that causes a signal to exceed a threshold is clearly a random event. Wikipedia defines the probability of a random event as the *relative frequency of occurrence* of an experiment's outcome, when repeating the experiment.

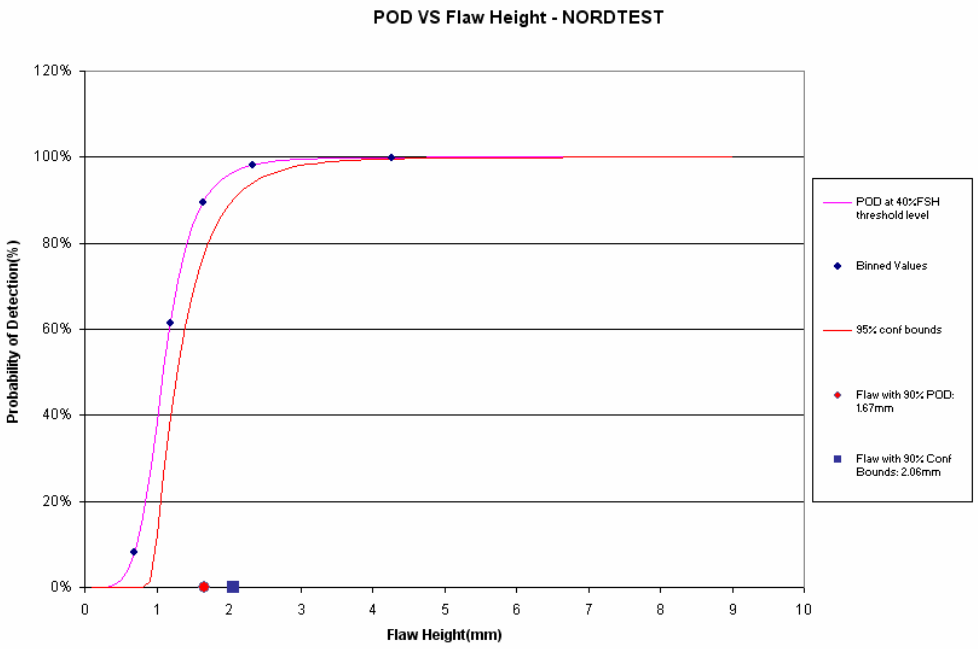
Repeating the "experiment" would be equivalent to repeating the test that was run to generate the scatter plot in Figure 1. There would be a good probability that the specific amplitude

responses from each flaw would not be the same as indicated in the graph. The flaws themselves would not have changed but other parameters could change resulting in smaller or larger amplitudes. It is most "probable" that the flaw causing the 1200% signal will always produce a response well above the 40% threshold but those at 35% to 45% on the original test may not always produce signals that are on the same side of the 40% threshold. This means that sometimes such a flaw is acceptable and other times it will be unacceptable.

The pipeline construction industry was one of the first to codify the probabilistic nature of NDT. O. Forli [3] had written on these concepts for some time prior to incorporating the ideas into the DNV off-shore construction code, OS F101 in the year 2000 [4]. Appendix E H300 of OS F101 stated that "The detection ability of an AUT system shall be deemed sufficient if the probability of detecting a defect of the smallest allowable height determined during an Engineering Critical Assessment…is 90% at a 95% confidence level."

When the process of probability determination for a flaw to produce a signal over the evaluation threshold is run against the flaw size the typical "S" curve usually results. Figure 3 indicates the Probability of Detection (PoD) for the amplitudes and flaw heights in Figure 1.

Figure 3 POD Curve based on Amplitude over 40% FSH Threshold



In 1997 Forli [3] wrote that the idea that a flaw had a probability of being rejected could be equated to its probability of detection. This was rationalised by the fact that any flaw that was to be rejected first required that it be of sufficient amplitude to be evaluated (detected).

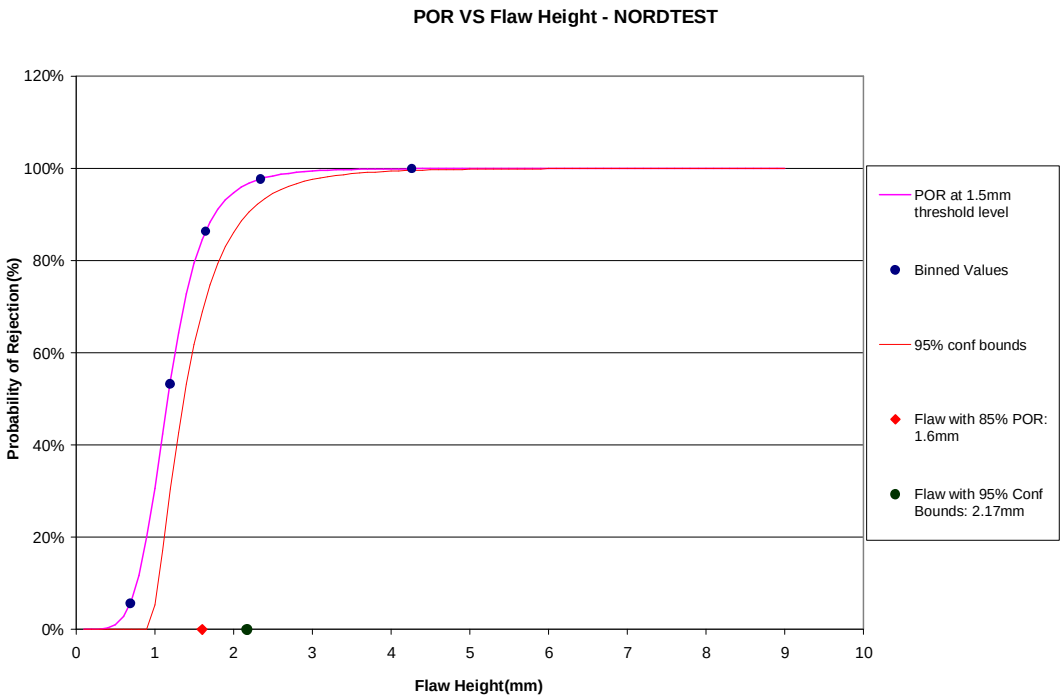
In the 2007 Edition of the DNV OS F101, the concept of Probability of Rejection (PoR) was revisited and it no longer holds the status of equivalent to PoD.

Parallel to the requirement to "detect" flaws, fitness-for-purpose acceptance criteria impose a requirement to size flaws. When the sizing ability of the ultrasonic system is incorporated into the assessment process new variables are introduced. No longer is the system tied to just a basic "maximum amplitude". Now other factors may be used. Forward or backward scattered tip diffraction sizing and zonal apportioning of signals can be used to modify the operator's estimate of a flaw size. Although this does not usually reduce the number of flaws being evaluated it can reduce the perceived severity of the assessed flaw size. It should be noted that the operator is not simply rejecting flaws based on amplitude when an ECA acceptance criteria is generated. Instead, the operator uses a graph produced by the engineers to compare the sized indication to an allowed size.

The sizing process is incorporated in the 2007 edition of DNV OS F101 Appendix E [5]. H303 now states: "*The detection criterion of H301 and the undersizing tolerance specified in H302 may be combined into one rejection criterion; There shall be more than 85% probability of rejecting a defect which is not acceptable according to the ECA determined criteria. This shall be shown at 95% confidence level.*"

This process is similar to PoD determination; however, instead of using the amplitude threshold directly the process compares the flaw size to the estimated flaw size using the critical flaw size as the threshold. The resulting graph is a PoR plot as indicated in Figure 4.

Figure 4 PoR Curve based on Sized Flaw for 1.5mm flaw height threshold



The plot in Figure 4 indicates that if the AUT system rejects all flaws calculated (by the AUT sizing technique) over 1.5mm in vertical extent then there will be an 85% probability (with 95% confidence) that no flaw greater than 2.17mm will go un-rejected.

Sizing and Amplitude

The notion that amplitude has some direct and independent link to flaw height is not substantiated in weld testing. The method of assessment known as "â versus a" illustrates this when the log-log plot is made of amplitude versus destructive test size as compared to the log-log plot of the UT "estimated" size versus the destructive test size. These plots are illustrated in Figure 5 for the same data set as was used in the previous Figures. The Height versus Amplitude is on the left and Height versus Height on the right (i.e. PoD compared to PoR).

Figure 5 Comparing log-log plots in "â versus a" calculation of PoD and PoR

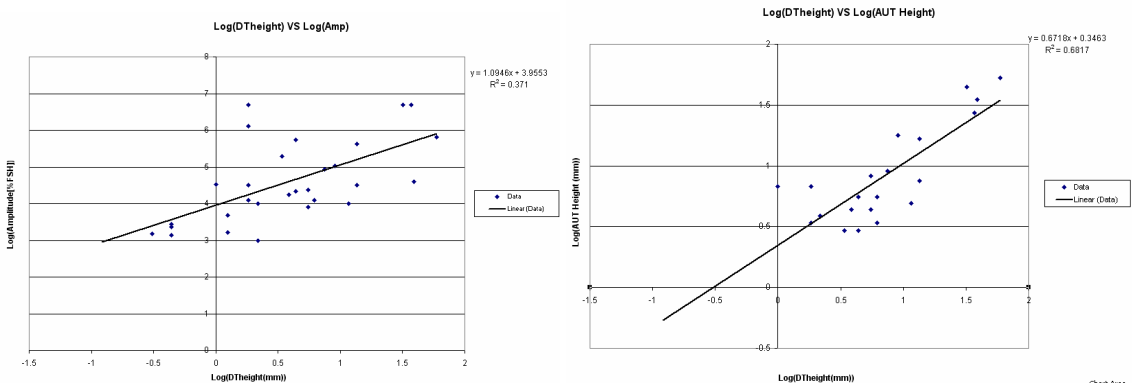


Figure 5 illustrates that the presumption of linear relationship of amplitude to flaw size is not well supported but when the amplitudes are "corrected" to calculate an estimated size the relationship becomes much more linear as the points are less scattered. The use of the â versus a model seems more appropriate when a size estimate is used instead of the traditional percentage screen-height amplitude.

The cumulative probability function for the â versus a model is

- $POD(a) = 1 - Q\left(\frac{\log a - \mu}{\sigma}\right)$ where Q is the standard normal survivor function,

  a is the flaw size, μ is the mean and □ the standard deviation.

This equation is almost identical to the hit/miss model. However, the â versus a model has as its underlying assumption that log(â) and log(a) are linearly related.

SUMMARY

The authors have developed a template using a spreadsheet data entry format that permits the calculation of PoD and PoR to comply with the requirements of DNV OS F101 2007. The probabilities are plotted with the confidence curves and the associated statistics for the data sets are also provided in graphed formats.
Report outputs in the form of tables and graphs provide useful understanding of the tolerances that can be expected with the equipment and sensitivity settings of any inspection system. These tolerances can be used to match the system setup to the ECA-based acceptance criteria being used on a project.
For more information on the Status3 template and the calculations it produces contact M. Matheson mmatheson@eclipsescientific.com

REFERENCES

1. Krautkrämer, J., Fehlergrössenermittlung mit Ultraschall. Arch.f.d. Eisenhuttenwesen 30 (1959), pp 693-703

2. Ermolov, I.N., The reflection of ultrasonic waves from targets of simple geometry", Nondestructive Testing 5 (1972), pp 87-91

3. Forli, O., http://www.ndt.net/article/0498/forli/forli.htm, How to develop acceptance criteria for pipeline girth weld defects, Det Norske Veritas, Oslo (N) European-American Workshop Determination of Reliability and Validation Methods of NDE, Berlin - June 18-20, 1997

4. DNV OS-F101, "Submarine Pipeline Systems, Appendix E, Automated Ultrasonic Girth weld Testing, January 2000.

5. DNV OS-F101, "Submarine Pipeline Systems, Appendix E, Automated Ultrasonic Girth weld testing, October 2007.