



PoD & PoS Study

Bobby Freer – P141566

Lectures – Ashely Pullen / Peter Charlton

Submission: 12/12/2017

Abstract

This paper investigated the Non-Destructive Testing (NDT) probability of detection and probability of sizing methods used throughout NDT industry. Ten operators were chosen at random to complete a detection and sizing exercise, from their recorded data and statistical analysis we found that operator seven was the most successful due to the high PoD results where 90% confidence was achieved at 0.85mm. For the PoS study it was found that operator eight was more accurate due to a higher R^2 value and low confidence bounds. After it had been narrowed down to a possibility of two operators for the most accurate PoS recordings, it was then determined that operator eight had a higher degree of accuracy after scoring an average of 0.9892 once we had combined both R^2 values and slope from the graphs, with the target number being 1.

1.Contents

1.Contents.....	2
2. List of Tables.....	3
3.Table of Figures.....	3
4. Introduction	4
5. Literature Review	6
5.1 NDT Statistics	6
5.2 Mean	6
4.3 Standard Deviation	7
5.4 Confidence Limits.....	9
5.2 PoD	11
5.2.1 PoD Case Studies.....	14
5.2.2 PoD Limitations	15
5.3 PoS	17
6. Analysis	19
6.1 Probability of Detection	19
6.1.1 PoD Operator Analysis	21
6.1.2 PoD Operator Comparison.....	26
6.2 Probability of Sizing.....	28
6.2.1 PoS Operator Analysis.....	29
6.2.2 R-Squared Values Comparison.....	32
6.2.3 $m \times C$ Comparison	34
6.2.4 Combining PoS Data.....	35
6.2.5 PoS Operators Overview	37
7. Discussion.....	40
8. Conclusion.....	42
9. Bibliography & References.....	43
10. Appendix	44

2. List of Tables

Table 1 Ultrasonic mean example.....	7
Table 2 Ultrasonic mean example comparison.....	7
Table 3 Standard deviation ultrasonic comparison	8
Table 4 Operator overview mean and standard deviations	37

3. Table of Figures

Figure 1 Mean calculation example	6
Figure 2 Standard deviation formulae	8
Figure 3 Standard deviation variables	8
Figure 4 Confidence limits example.....	9
Figure 5 Pod curve example.....	13
Figure 6 Pod hit/miss formulae.....	13
Figure 7 Pod hit/miss and signal comparison	16
Figure 8 PoS example.....	18
Figure 9 Operator PoD defect hit/miss recordings	19
Figure 10 Operator seven PoD.....	21
Figure 11 Operator eight PoD	22
Figure 12 Operator four PoD	23
Figure 13 Operator one PoD	24
Figure 14 Operator nine PoD	25
Figure 15 PoD comparison	26
Figure 16 PoD size of Operator defect recordings.....	27
Figure 17 PoS Operator two.....	29
Figure 18 PoS operator eight	30
Figure 19 Operator ten PoS	31
Figure 20 PoS R-squared comparison	33
Figure 21 PoS y intercept comparison	34
Figure 22 PoS Combining values Comparison.....	36
Figure 23 PoS operators mean & standard deviations	38
Figure 24 PoS y-intercept calibration issues	39

4. Introduction

NDT processes and procedures can be ineffective unless the correct method is applied to the task in hand. We could define what could be the optimum process and procedure for a given task by using statistical analysis on experimental trials. By optimising the methods of testing with statistical analysis we will be able to get more accuracy from readings, and improve the chances of an operator finding a flaw.

As NDT engineers we can use basic methods of statistics to analyse our data such as the mean, standard deviation and, confidence limits, these analytical methods will indicate how consistent tests may have been but to get more indication on the testing process and procedure we can use PoD (Probability of Detection). PoD can be defined as the probability that using a given inspection procedure/system/sensor an inspector will detect a flaw that is present in a component (B. Raj, 2007). As well as PoD giving information on flaw detection we can also analyse operator capabilities of defining the size of a known defect using Probability of Sizing (PoS).

The ability to check how accurately an operator can define a defects dimensions, will allow designers to take into account the size of defects which could be missed in an inspection when they are calculating fracture mechanics, this then allows them to calculate a safety factor for the component.

In this report two sets of data from an ultrasonic inspection have been analysed using PoD and PoS.

In total twenty different operators provided data, ten for the PoS and ten for the PoD. The aim was to be able to determine which operator recorded the most accurate sets of data, then consider why have they recorded this data, for reasons such as machine calibration and consistency of sampling.

It was found for the Pod that the operator who recorded the most consistent data did not find the smallest defect out of the group of the operators, and was proven that the group of operators struggled to record any defects under 0.25mm. In the PoS study it was found that there was an obvious visual separation that could be determined between operators, but to justify who recorded

the most accurate data we would need to use deeper analyses of trendline correlation and some calculation.

5. Literature Review

5.1 NDT Statistics

In the Non-destructive Testing industry, we will be recording different types of data through testing and analysis on both equipment used and procedures followed to try to determine their capabilities of finding a defect. To ensure that we are getting what we need from the equipment and procedures statistical analysis can be carried out as a means of comparison. Three basic forms of statistical comparison that can be used for NDT are listed below;

5.2 Mean

The mean is a type of average which is the most common statistic of choice when frequency distributions are fairly symmetrical. There will be two types of mean, sample and population mean. The difference between the two being that population mean would indicate that every physical piece of data available has been used for the calculation, whereas sample mean states that the calculation only uses samples of the population. In any event, the formulas state the same thing, the only difference is the procedure for calculating it. Sample and population mean equations are displayed below (L.Phillips, 1973).

Sample Mean	Population Mean
$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{N}$

where $\sum x$ is sum of all data values

N is number of data items in population

n is number of data items in sample

Figure 1 Mean calculation example

The most commonly used form of mean is the sample mean as it is not always possible to gain the data needed to calculate a population mean. In terms of NDT, the mean is a useful tool for data analysis as it will give us a more general description needed rather than data at specific points.

For example, in an ultrasonic test to measure attenuation in different materials, there was variance in the readings due to factors such as material impurities and defects. A calculated mean value gave an indication of how the sample piece behaves as a whole, rather than an individual point as the table below displays.

Table 1 Ultrasonic mean example

POSITION	Attenuation
1	-1.83
2	-1.10
3	-1.53
4	-1.42
5	-1.28
6	-1.69
Mean	
-1.48 Decibels	

All mean calculations in this report involving a mean value will be assumed to be a sample mean as it is not stated that a population of data has been recorded.

4.3 Standard Deviation

Standard deviation is a measure of variability in a set of data. It states how much each recording varies from the mean value of the data set. For example, if we look at the ultrasonic test on attenuation again we have two tests run on the same sample below which both have the same mean, but the sets of data are very different from one another. Therefore, we could have two very different sets of data which still give the same mean. We need a standard deviation to define the sets of data.

Table 2 Ultrasonic mean example comparison

Test One		Test Two	
POSITION	Attenuation	POSITION	Attenuation
1	-1.83	1	-1.90
2	-1.10	2	-0.03
3	-1.53	3	-0.3
4	-1.42	4	-3.92
5	-1.28	5	-0.08
6	-1.69	6	-2.89
Mean		Mean	
-1.48 Decibels		-1.48 Decibels	

Standard deviation can be defined by the two equations below. Again, here like the mean, we will have a sample and a population standard deviation.

Sample Standard Deviation: $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

Population Standard Deviation: $\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$

Figure 2 Standard deviation formulae

The fundamentals of the two equations are the same;

X = All values in data set
 $\bar{X}$ = Mean of all values in data set
 n = Number of data recordings
 $n - 1$ = Number of recordings -1
 (Sample recordings)

Figure 3 Standard deviation variables

The only difference being $n - 1$ indicates we are using sample standard deviation. The reason for this $n - 1$ being there is if you take a sample mean you know it will not be the same as what a population mean will be, it will be a little bit off. Likewise, if you take the standard deviation of a sample mean it will also be a little bit off from what the population standard deviation would be. We don't want the sample deviation always to be slightly too small underestimating the actual deviation, so if we add the - 1 (making the denominator smaller) to the division it makes up for the fact that small samples have less spread than bigger samples or a population recording (L.Phillips, 1973).

Table 3 Standard deviation ultrasonic comparison

Test One - Standard Deviation	Test Two - Standard deviation
0.27	1.65

If we now include a standard deviation for the two sets of results we can see how they vary from the mean. We can see that Test Two deviates much more from the mean than Test One does, this

information could now be plotted graphically to represent the deviation of results, but what is important is that it gives another analysis tool to identify data behaviour which the mean would not.

5.4 Confidence Limits

The purpose of taking a random sample from a population, and then calculating the mean for example, is to approximate the mean of the population. How well the sample statistic estimates the underlying population value is always an issue. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest (itl.nist.gov, 2013).

Confidence limits are constructed at set parameters by the user such as 95%. This means that if the same random sample from the population is sampled on numerous occasions all samples should sit within the 95% confidence limits for 95 % of the cases.

The confidence interval will form a bracket above and below the point of interest (i.e. the mean), known as the upper and lower limits as the figure below displays.

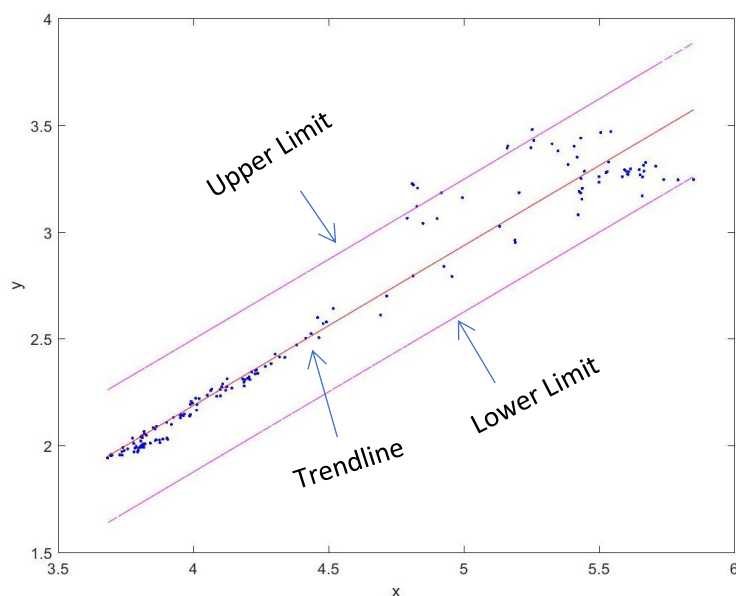


Figure 4 Confidence limits example

We can calculate confidence limits using values from the equation of a straight line (which could be found the trendline in this case) and the standard deviation.

$$\pm 95\% \text{ Limit} = mx + (C \pm (1.96 * \sigma))$$

(1.96 comes from fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean (itl.nist.gov, 2013)).

Confidence will vary based on the data which it is created from, a wider spread of data results will create a wider 95% confidence band, similarly a tight set of data will produce a narrow confidence band.

5.2 PoD

Some of the most important questions to be answered when looking to use NDT for a specific task are, what inspection methods are applicable? how small a flaw can they detect? And what type of process control is needed to maintain this level of performance? The wide range of NDT processes capabilities can depend on factors such as calibration in an automated eddy current process, or operator skill in a liquid penetrant. The variances in NDT processes can be addressed using PoD concepts (B. Raj, 2007).

PoD (probability of detection) is a method to validate testing equipment and methods in relation to a defect size. It was developed by NASA in the late 60's as there was a necessity to validate procedures and capabilities of procedures and NDT equipment. An example of this variance that caused the need for PoD is in a paper by WH Lewis, (Lewis W H, 1978) it is described how he had 60 air force inspectors, use the same surface eddy-current technique to inspect 41 known cracks around countersunk fastened holes in a 1.5m length of a wing box, the results were that none of the cracks were detected 100% of the time, and different cracks with the same size had variation in detection percentages. Case stories like in WH Lewis paper can be found commonly in NDT literature, this is why there was a need to measure the variance in NDT process capabilities. (Although the main reason for NASA in developing PoD was more to find out the largest possible defect that could be missed for various NDT methods, and use this figure in consideration for fracture mechanics and design). Now many industries have adopted PoD as their standard NDT reliability method.

There are no universally fixed procedures for PoD, as procedures are specific for each individual task. The importance of understanding PoD is crucial for the validity of situation they are being applied in, unless the pure concepts of PoD are understood it will be difficult to apply it to the situation in hand as every case is specific. PoD will depend on a number of physical parameters such as material, thickness of component, flaw type, geometry etc which must be appreciated (Jacobi, 2006).

PoD can be found by experimentation on the number the number of defects found in parts that contain a known number of flaws, i.e. if 100 defective parts are inspected and 95 are found to have defects we can say that the Probability of detecting a defect is 95% (B. Raj, 2007). PoD results will fall somewhere between 0 (0% chance of detecting a defect) and 1 (100% chance of not detecting a defect), i.e. if a defect is found a value of 1 is recorded and no defect records 0. By assigning POD values to a procedure, it is possible to compare the effectiveness of different inspection techniques and equipment for a certain task. Alternatively, PoD values can be used as a target for procedures to be shown to be able to achieve (www.bindt.org, 2013).

This experimental process mentioned of whether a flaw has been found in a sample with known defects, will categorise results into four possible outcomes (www.ndt.net-POD, 2006). It will essentially be described as hit and miss data.

1. An item is flawed, and the NDT method detects it (True Positive)
2. No flaw exists, and the NDT method indicates a flaw present (False Positive)
3. An item is flawed, and the NDT method does not detect it (False Negative)
4. No flaw exists, and the NDT method has no indication of a flaw (True Negative)

The tool most commonly used for PoD description is the PoD curve, as displayed in the graph on the next page taken from an NDT.net (www.ndt.net-POD, 2006). Here we can see from the graph that there is a 30% chance of detecting a flaw 1mm high, and an 80 % chance of detecting a defect 2.2mm high. The pink curve below the PoD curve displays a confidence limit.

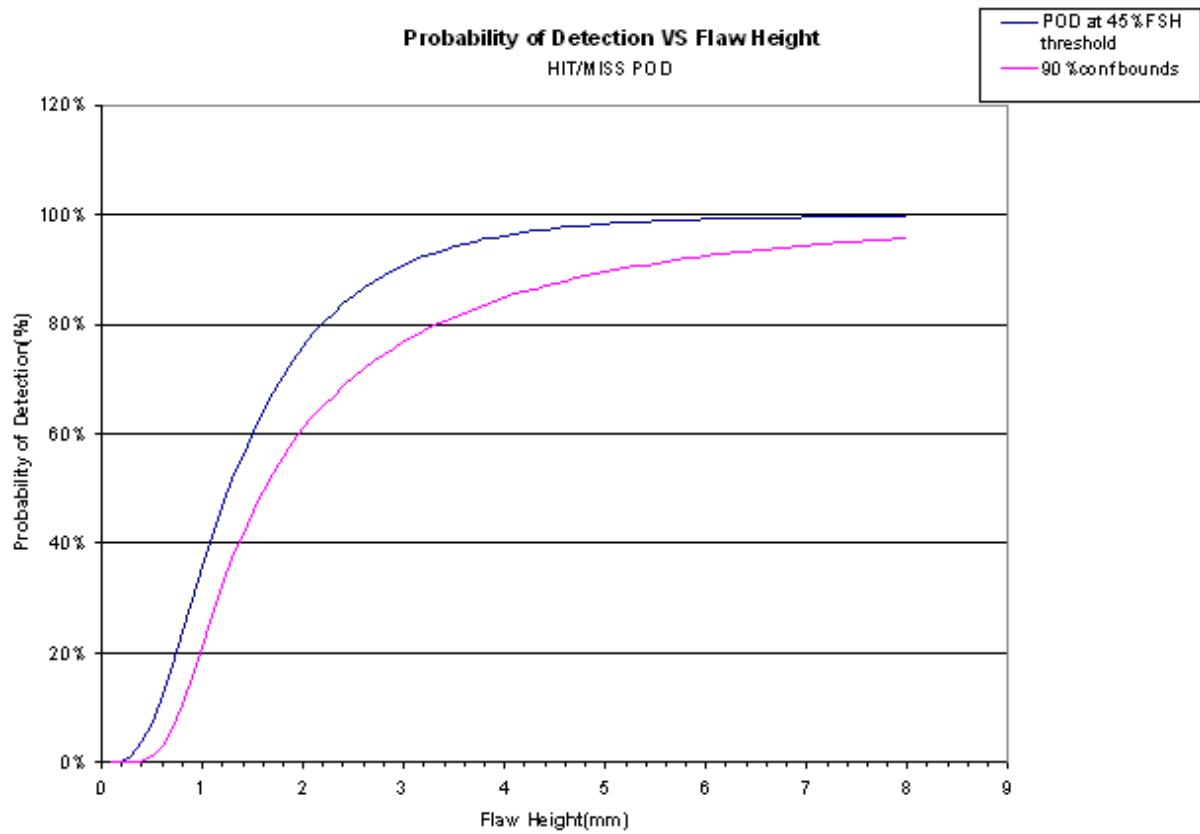


Figure 5 Pod curve example

As mentioned PoD is specific to individual cases and techniques, such as if we were using signal response data to build the curve rather than hit/miss we would use a specific model for each set of data (Jacobi, 2006). This report will concentrate on a PoD model using hit/miss data, which is represented by the equation below.

$$POD(\alpha) = \frac{e^{\frac{\pi}{\sqrt{3}}\left(\frac{\ln \alpha + m}{\sigma}\right)}}{1 + e^{\frac{\pi}{\sqrt{3}}\left(\frac{\ln \alpha + m}{\sigma}\right)}}$$

α = Defect Dimension

σ = Standard Deviation

m = Mean

Figure 6 Pod hit/miss formulae

The standard deviation and mean will be calculated from the hit/miss data, where the defect dimension will be measured with equipment. Using this PoD model, we can create the curve.

5.2.1 PoD Case Studies

PoD has improved NDT results through many industries such as offshore, rail and nuclear to name just a select few (Jacobi, 2006).

Offshore

Underwater PoD trials at University College London in the 1990's considered the detection of fatigue cracks in offshore tubular joints (Jacobi, 2006). The results of the trials were used to compare the flaw detection capabilities of Magnetic Particle Inspection with a number of eddy current NDT techniques as well ultrasonic NDT techniques using creeping waves. For all techniques considered, a POD confidence limit was achieved for cracks with typical lengths ≥ 100 mm (Jacobi, 2006).

Rail

PoD is commonly used in the rail industry to quantify reliability and to optimise the inspection methods. POD has improved estimates for ultrasonic axle inspection, PoD estimates have also been produced for the improved NDT methods and for new designs utilising hollow axles.

A simulation model utilising real A-Scan data and data from real flaws was used to produce POD curves for far-end and near end axle inspection. This enables specific PoD curves to be produced for individual axle designs and geometries (Jacobi, 2006).

Nuclear

The ultrasonic NDT procedures used in the trials of steel components in thick walled pressure vessels for the nuclear industry during the mid-1970's was applied too rigidly and did not allow the signal responses from large planar flaws to be evaluated properly. This caused low PoD readings for quite large flaws. This is an example of a poorly designed NDT procedure.

Here PISC (The Programme for the Inspection of Steel Components) (PISC-I, 1979) was bought in and allowed their testers to use their preferred NDT method which proved to give much higher PoD for the same flaws. This improvement came from changing NDT procedure where the improvements and changes were recognised with PoD.

5.2.2 PoD Limitations

On some occasions the behaviour of the PoD curves may appear to give illogical results. The models used to determine the PoD curve for hit/miss data and the signal response data are quite different.

The hit/miss and signal response functions are very similar in their same statistical parameters, and if the wrong one is used, the values will be slightly out of what is expected (see figure 7).

It may be the case that other modeling approaches need to be considered if the PoD is giving analysis which makes little sense. However, it will be useful to carry out some checks to see if there is something specific about the data in order to decide what action to take.

For hit/miss data it has been found that the PoD function can sometimes decrease as flaw size increases. This is usually because the NDT experiment was badly designed, and it would then require a repeat some of the trials with better designed NDT procedures to get more accurate results. In Section 2.2.1 the nuclear test describes how a poorly designed NDT experiment led to unexpected results, the solution was to run more trials with a better NDT procedure. Here the capability of POD was not able to be utilised due to a poorly designed experiment.

When recording hit/miss data over a defect, the region of where the flaw hits and flaw misses should not be distinct, there has to be a good overlap otherwise the curve that fits the hit/miss model will not produce a valid solution. This usually means more data is required in the region between the smallest and the largest sizes of the flaw (Jacobi, 2006).

Generally, we are stating here that for accurate PoD which can be trusted a well thought out testing procedure must be carried out diligently by the operator.

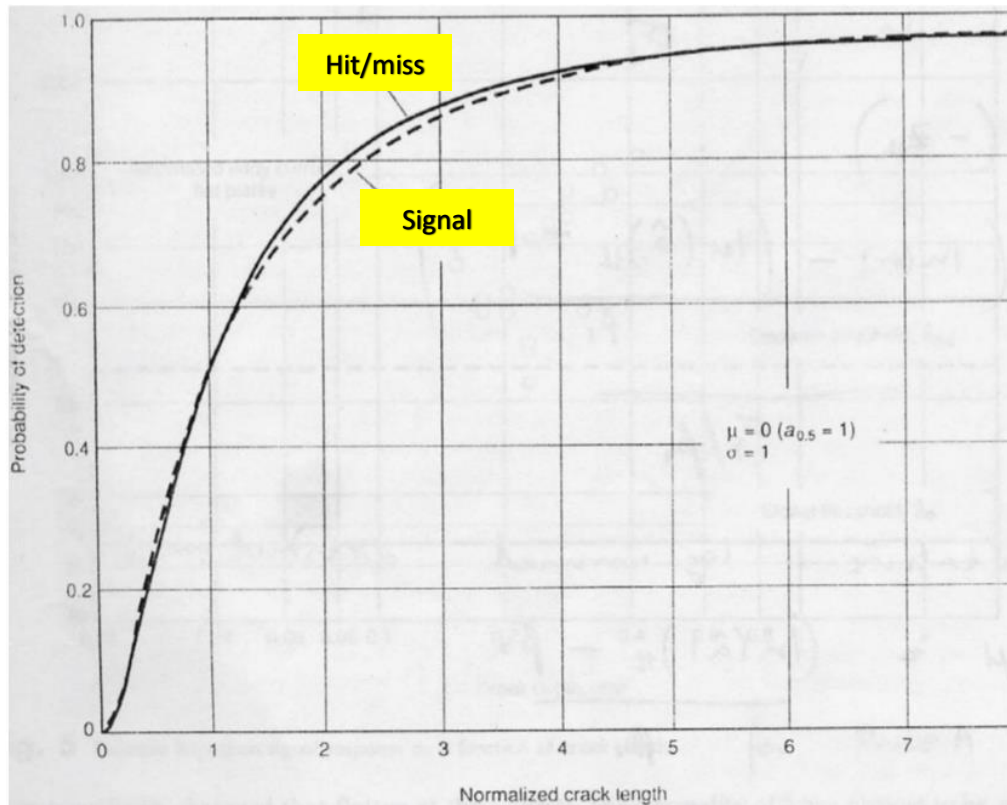


Figure 7 PoD hit/miss and signal comparison

The figure above shows how close the relationship is between using a PoD signal model and a PoD hit/miss model. If the wrong model is used for analysis the results may be so similar that it will be difficult for the engineer analysing to spot the mistake (Berens, 1997).

5.3 PoS

Probability of Sizing refers to how accurately an operator can size a defect. Here uncertainties will exist when the operator must provide dimensions on the flaw, different operators readings will vary plus/minus the actual size of the flaw, as is typical with all operator measurement. For example, if a group of 20 people were to use measure a block with a vernier calliper for *Length, Height, Width* results would differ due to how much pressure they apply to the calliper jaws, where they take the measurement from, skill level and if it is a mechanical, not a digital vernier gauge possible rounding errors (www.ndt.net-POD, 2006).

PoS is important as sizing inconsistency will relate back to fracture mechanics already calculated in design. When an NDT operator is asked to make a judgment about the size of the flaw the result is then compared to an allowable dimension chart, operator under-sizing could ignore a potential failure and oversizing could cause concern to a within tolerance component. The graph below has an Ideal Sizing line, this would have been calculated through destructive testing where the recordings can now be compared to what is an allowable deviation. In the case of where a known defect sample is being used we can create a True Size line and compare the operator's readings to this True Size, this will give assessment on how well the equipment can be used by the operator.

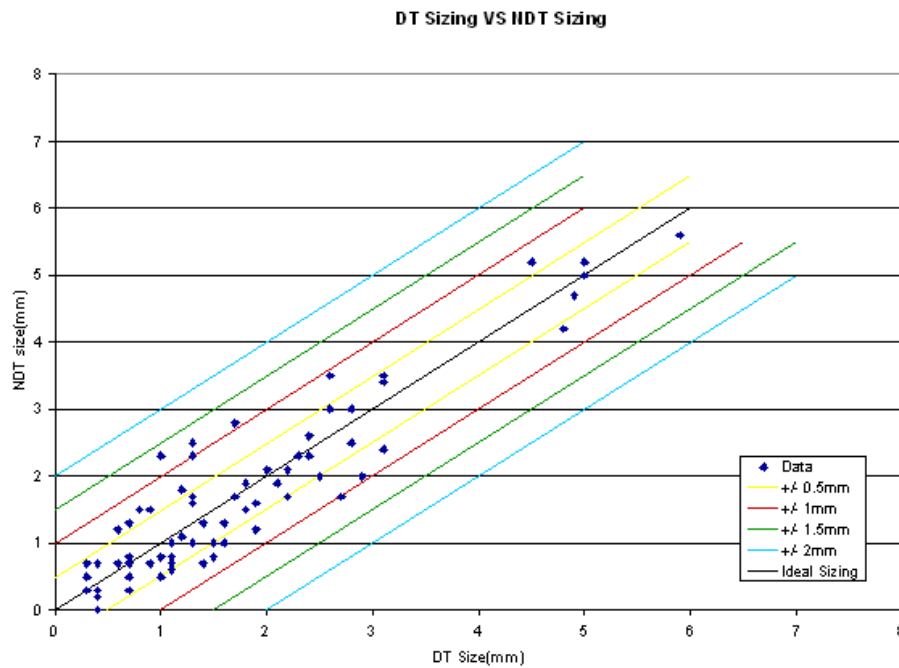


Figure 8 PoS example

These graphs are created by comparing the true size against the operators estimated readings. From these values the operators error can be calculated using the following;

$$\text{Operator Error} = \text{Operator Measurement} - \text{True Size}$$

This value is used to calculate the standard deviation of the operator's errors allowing for the 95% confidence limits to be determined, from using this value of Operator Error the equation found in section 2.3 can be used to calculate the standard deviation of results from the True Size.

6. Analysis

A PoD and PoS analysis will be performed on two sets of results taken from ultrasonic testing to assess the performance of the procedure. The statistical methods used here will be able to highlight possible problems with the equipment and the way the procedure is being carried out, as well as demonstrating reasons why certain operators have recorded more consistent data.

6.1 Probability of Detection

Defect parameters provided have been reduced to length and width values only. The samples include 28 defects in total, with ten technicians using the same procedure to inspect the samples.

As all operators were using the same equipment and same procedure, we will be looking at analysing operator accuracy rather than trying to evaluate different methods for the task.

Rather than try to give a detailed analysis on each operator, and over complicate data it was decided to concentrate on five of the ten operators. We would try to choose operators with wide variation to one another's readings, and also an operator who recorded data somewhere in the middle between the widest variation of results. To make this decision we will first look to see who recorded the most hits out of the 28 defects.

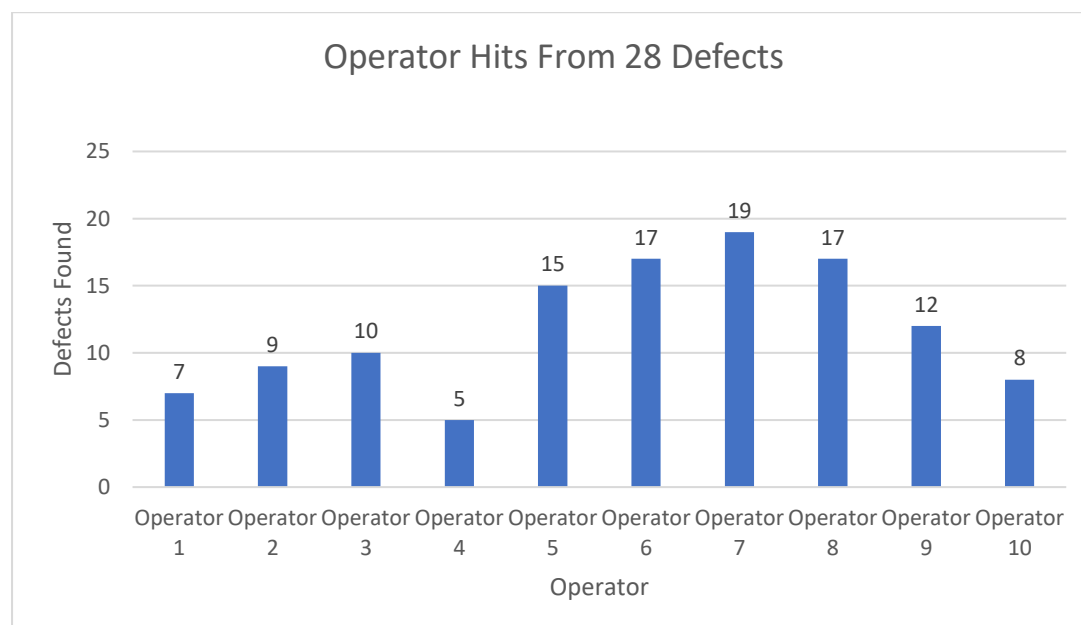


Figure 9 Operator PoD defect hit/miss recordings

We can see from the bar graph that operator four has recorded just five defects from the possible 28, whereas operator seven has recorded the most amount of hits where 19 were found. The average amount of hits from the operator's recordings was calculated to be 11.9;

$$\frac{119}{10} = 11.9 \text{ defects}$$

Therefore, on average we are looking at twelve hits (as we cannot have 0.9 of a hit), that will be recorded from the operators over the 28 defects. We will choose to analyse operators seven and eight, because they appear to have recorded the most number of defects, operator nine as the results produced represent the average, and operators four and one because they recorded the least amount of hits (the rest of the operators PoD can be found in appendix 1).

6.1.1 PoD Operator Analysis

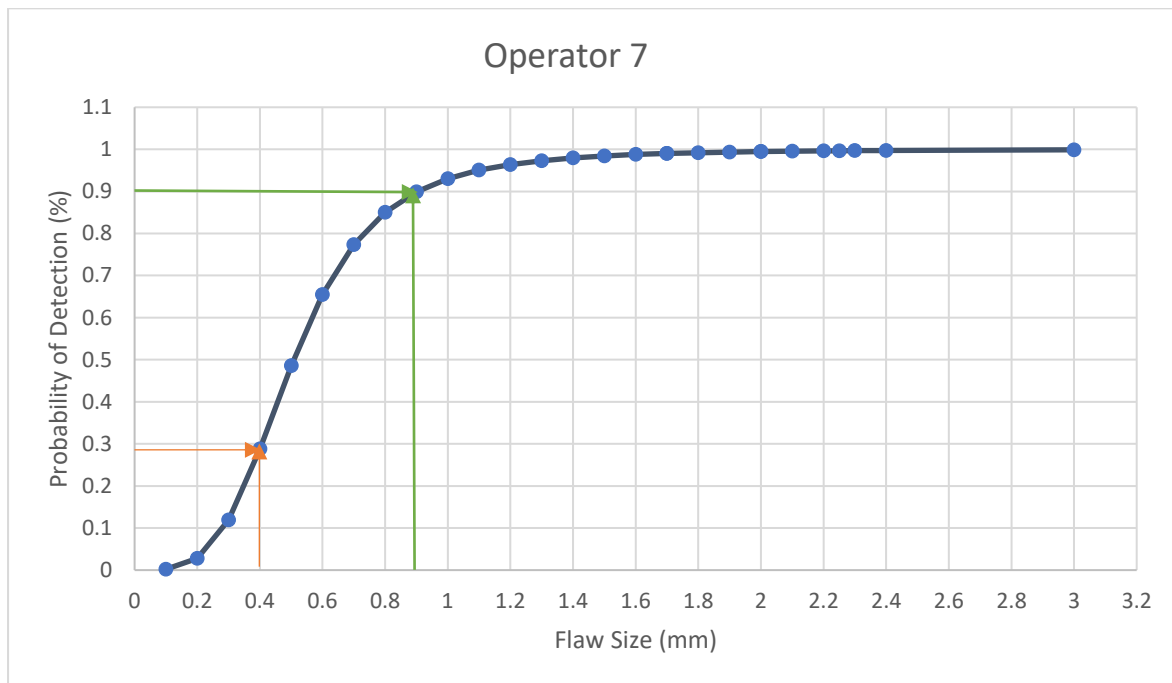


Figure 10 Operator seven PoD

We can see from the graph above that operator seven has a 90% confidence in finding any defect over 0.85mm. The slope of the graph moves moving quickly from 0.1mm up to 0.8mm in comparison to the other operators. Operator seven displays the best POD curve of the ten operators as he has a higher chance of detecting the smaller defects than the other nine operators, we can see that there is a 29% chance of detecting anything 0.4mm or over

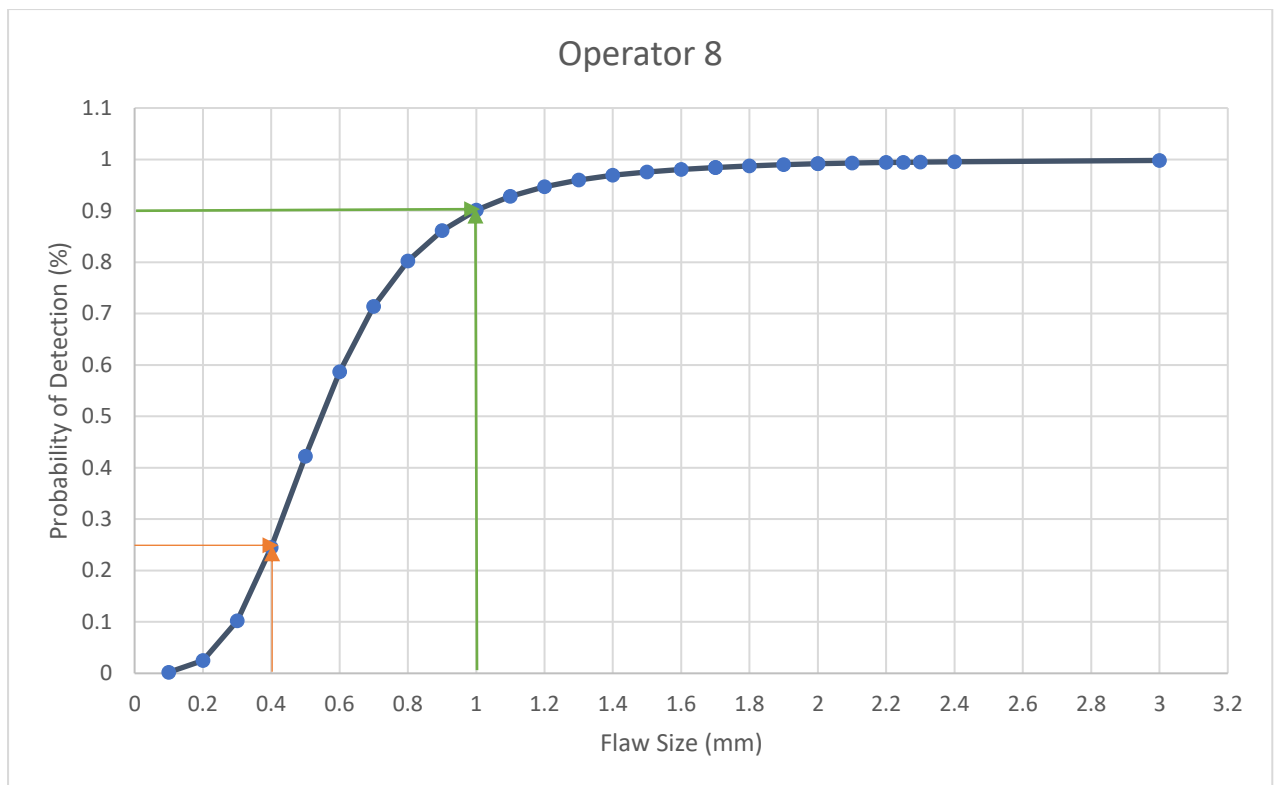


Figure 11 Operator eight PoD

In the graph above, operator eight's PoD curve is displayed. Operator eight found the most amount of defects after operator seven. If we concentrate here on small defects it appears that to find a defect of 0.4mm we would have a 25% chance, whereas operator seven would have a 29% chance. If we look when operator eight reaches 90% confidence we can see that for any defect over 1mm over 90% confidence will be achieved. It is in the smaller defect sizes where operators seven and eight stand out from the rest of the group, with only operator eight above coming close the operator's seven's capabilities overall. But it should be noted that operator eight found a defect at 0.9mm , whereas the smallest defect operator seven found was 1mm , after this though operator seven displayed more consistency and recorded all defects whereas operator eight did not.

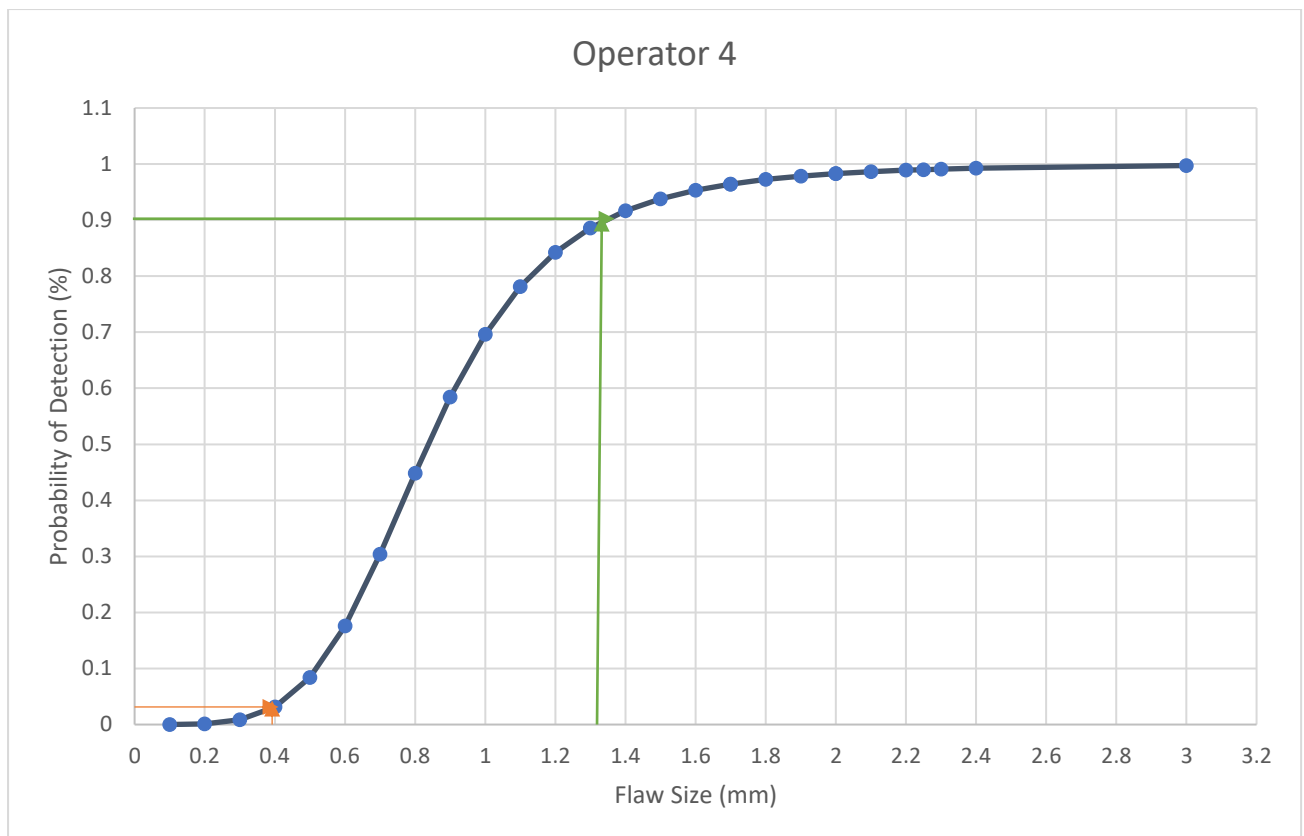


Figure 12 Operator four PoD

Operators four and one were deemed to be the two worst readings, in terms of these operators having the less likelihood of picking a small defect in comparison to the other eight operators. If we concentrate again on the possibility of detecting a 0.4mm flaw, operator four would only have a 0.035% chance of detecting this flaw. We can see here that there is almost 0% chance of detecting anything under 0.3mm . We also can see that 90% confidence (displayed by the green arrows) isn't achieved here until we encounter defect of 1.3mm and over, in comparison to operator seven achieving 90% confidence 0.85mm and over.

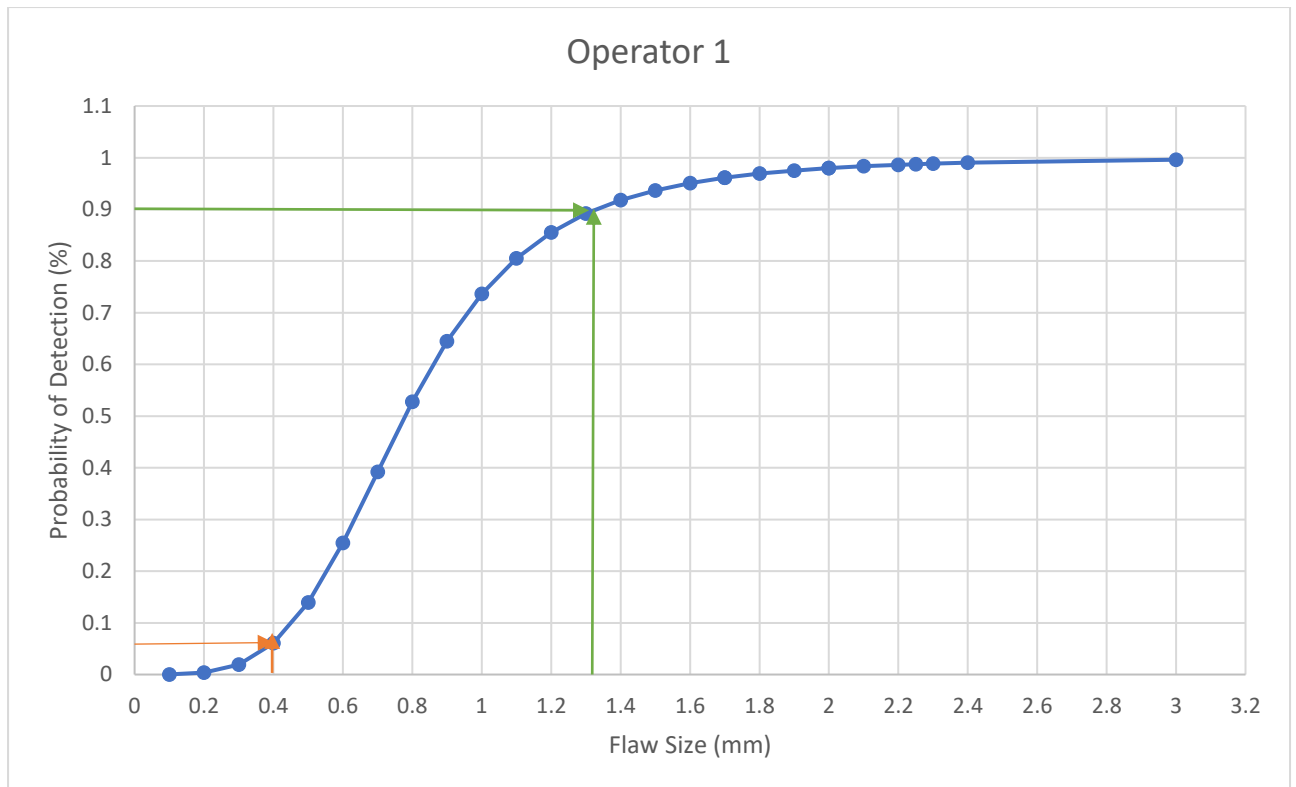


Figure 13 Operator one PoD

Operator one missed the second highest amount of defects in the sample, with only operator four missing more defect. Here we have almost 0% chance of detecting anything under 0.3mm.

Concentrating on 0.4mm again we would only have a 0.061% chance of detecting a defect of this size. And it will take until we start encountering defects of 1.35 until we gain 90% confidence from the readings.

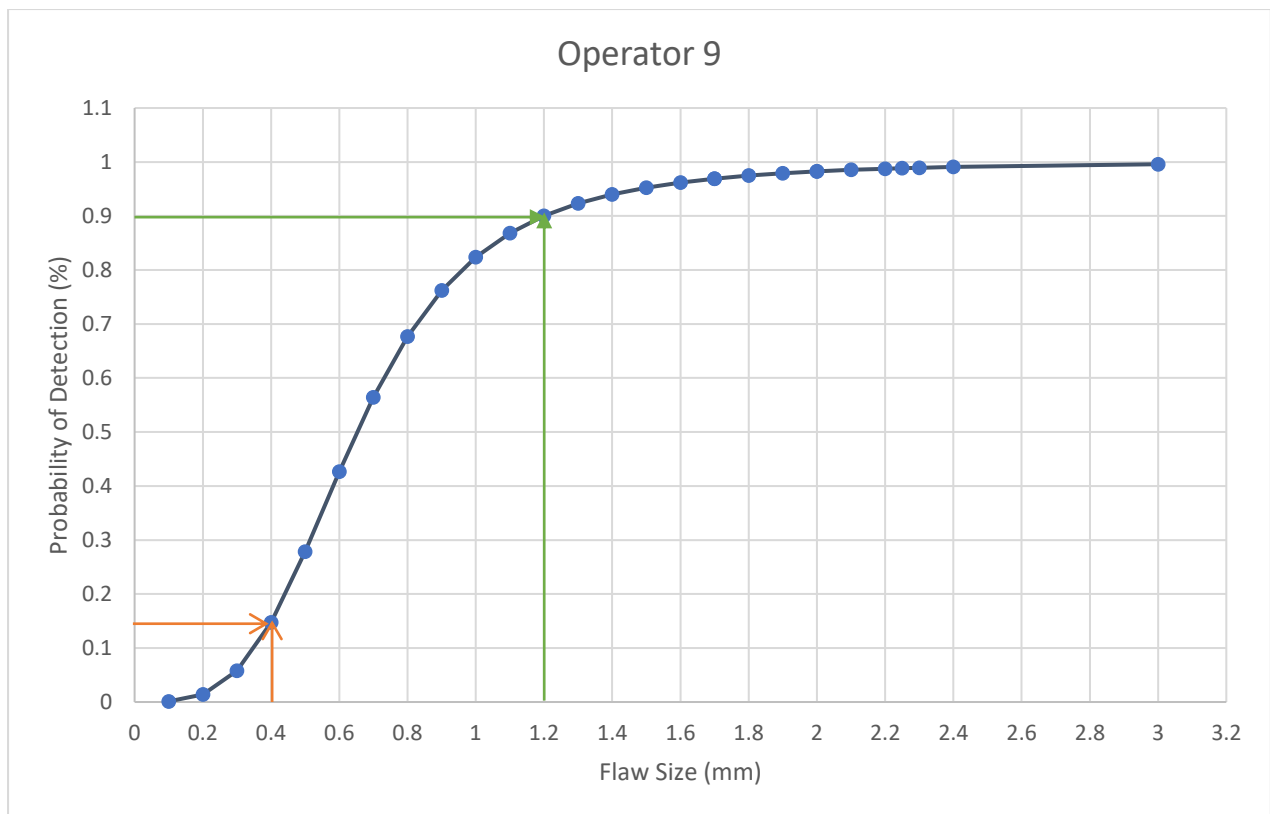


Figure 14 Operator nine PoD

values Operator nine's results represent which we calculated the average to be. This statistically represents what we could expect to find if when we have a group of different operators (and skill levels) testing the same sample. We can see that the probability here of detecting a 0.4mm is 15%. It takes until we encounter defects over 1.2mm to gain 90% confidence in the readings, which as expected sits somewhere in the middle of where 90% confidence was found in operators four and seven.

6.1.2 PoD Operator Comparison

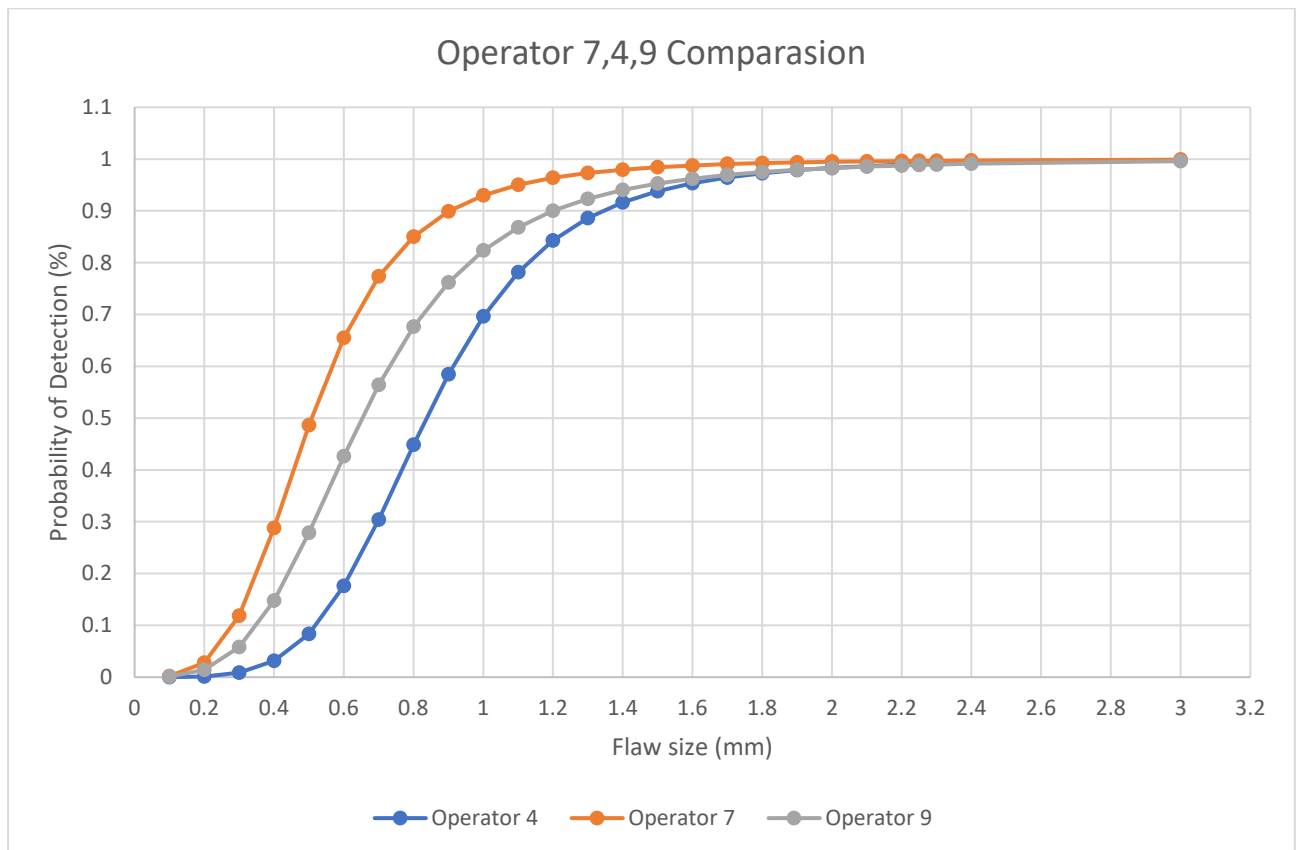


Figure 15 PoD comparison

In the graph above we can compare the plots of what we consider to be the best set of readings (operator seven), and the worst (operator four), and what we consider to be a set of data very close to the mean (operator nine). We can see that operator sevens curve has a greater rate of change than operators nine and four, meaning that the bigger the defects become operators seven will be exponentially more likely to find them in comparison to the other operators. As expected we can also see operator nines results sitting in the middle of the best and worst readings, therefore demonstrating the average.

None of the operators managed to record a defect under 0.9mm . Operator eight managed to record this, whereas the smallest defect operator four could record was 2.1mm . The graph on the following page displays this.

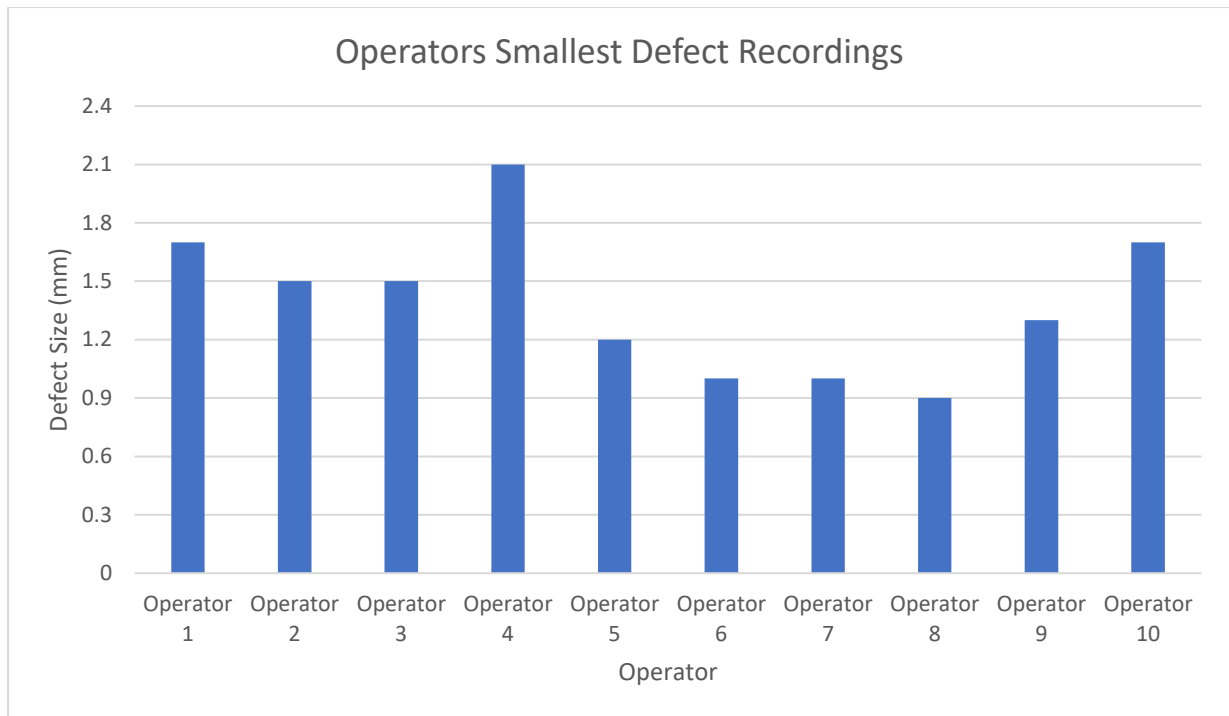


Figure 16 PoD size of Operator defect recordings

6.2 Probability of Sizing

Here we are measuring a new ultrasonic thickness gauging procedure in a sample with known defect sizes, to analyse how closely the operators can size the defects (appendix 2). There are 21 defect measurements in total, given to one decimal place. Ten technicians used the same equipment and same procedure to perform the measurements.

It was decided here that rather than thoroughly analyse all ten operators it would simplify analysis by just choosing a select group of operators to analyse.

We will try to focus on operators with the most variation to one another, then if there is correlation between certain groups of operators look to categorise them.

First, we will decide which operators to analyse. To make this decision, we will visually inspect the PoS graphical data, for all ten of the operators recorded in comparison to one another. It was decided from the visual inspection that operators eight and ten would be analysed because they display a more consistent set of data compared to the other operators, also operator two would be analysed as a worst-case scenario as operator, operator two appears to have the least consistent results as there is a large standard deviation resulting in wide confidence bands from the recordings (the remaining operator's results can be found in appendix 3).

6.2.1 PoS Operator Analysis

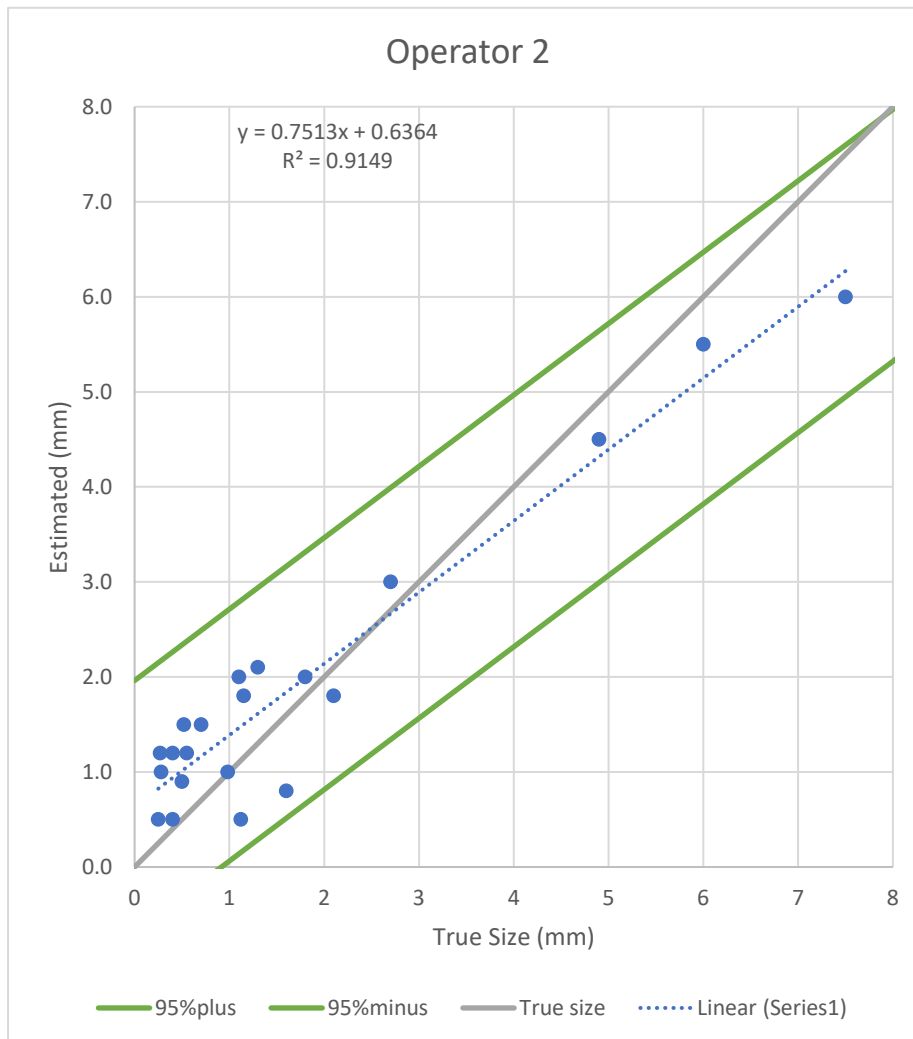


Figure 17 PoS Operator two

The graph above displays operator two, from visual inspection and checking the standard deviations of the operators we deemed this to be the least consistent set of data recordings. We can see that operator two on the whole is oversizing the measurements, displaying 6 readings below the true line, and 15 above the true line.

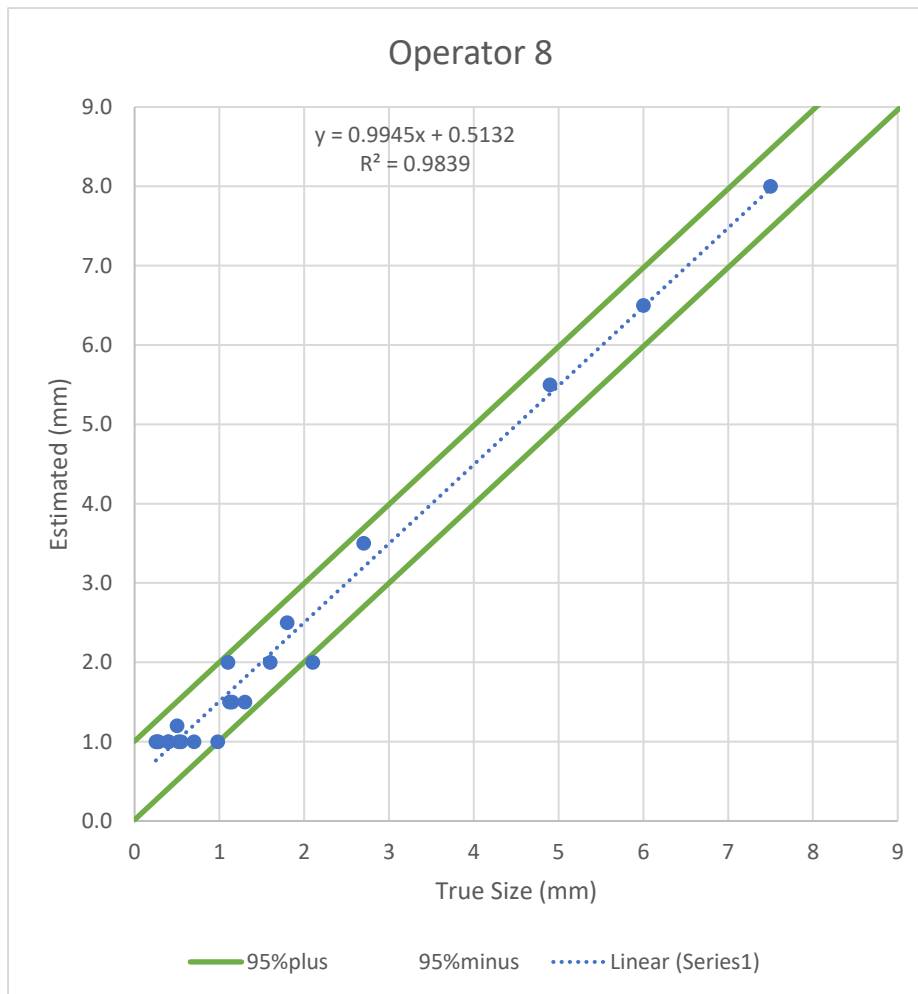


Figure 18 PoS operator eight

The graph above displays operator eight. We have considered operator eight one of the most consistent recordings of data, we can see the confidence bands are narrow indicating that we are looking at results with a small standard deviation, which means there must have been a degree of consistency in the recordings. A point of interest we can observe from the trendline is that the y intercept value of the trendline is 0.123 smaller than operator two, displaying that the equipment is likely in more reliable calibration than operator two.

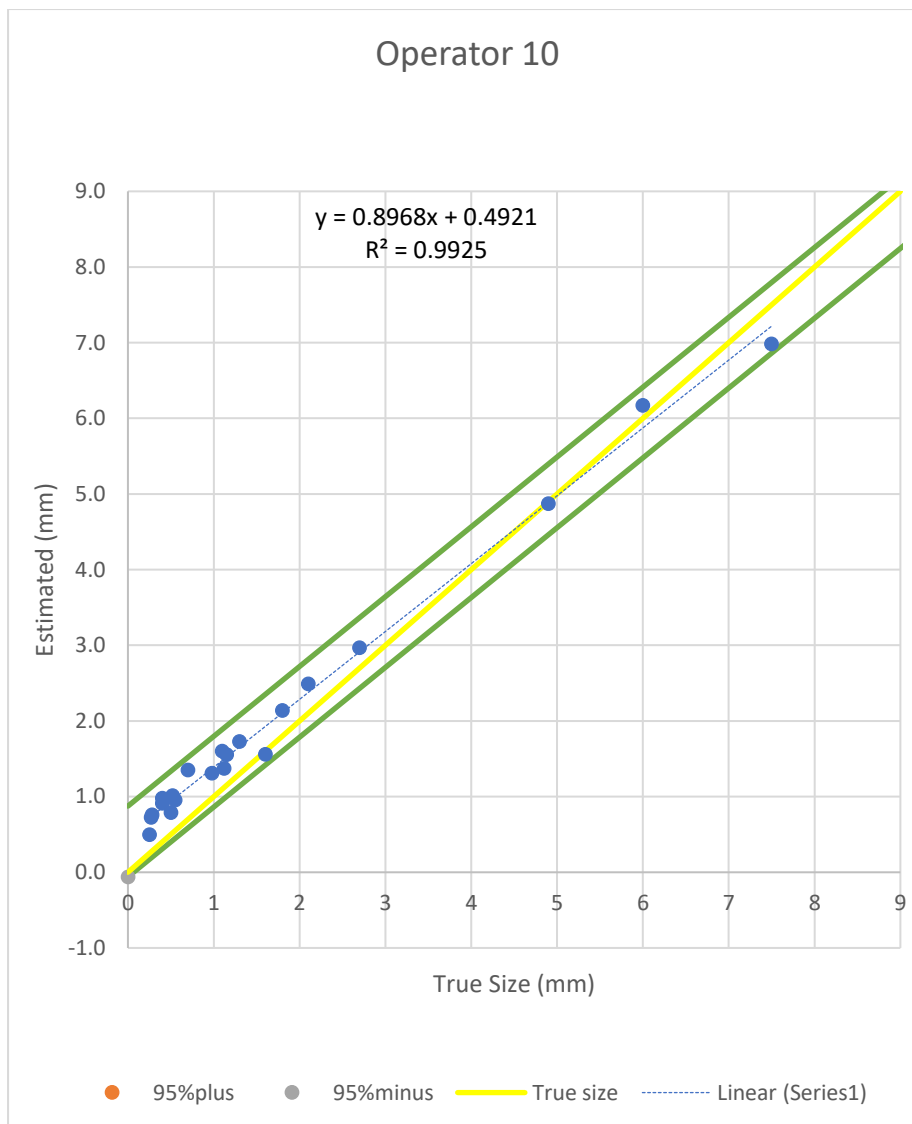


Figure 19 Operator ten PoS

Operator ten above displays confidence lines with a close bandwidth similar to operator eight. The data plots show oversizing for all but three values, this would indicate that rather than just being out of calibration, that the equipment may well be over calibrated slightly resulting in the majority of the points being above the true size.

6.2.2 R-Squared Values Comparison

We have narrowed it down to the point where we can graphically make the conclusion that operators eight and ten have recorded the most reliable data. We now must look to further than visual representation to split the two operators, and see who has recorded the most reliable data. If we now look at comparing the R^2 values of the three operators in question, we will be able to determine who has the most consistent set of data. Here we will look at operator two also, as a means to show a worst-case scenario in relation to what two consistent recordings give.

R^2 values demonstrate the predicative capabilities of a linear graph (people.duke.edu, 2001). The R^2 value is a statistical measure of how close the data points are fitted to the trend line, it is defined as the percentage of the response variable variation in a linear model. 0% would indicate that the model explains none of the variation, 100% would indicate that the model explains all variability of the data around its mean. So, we will be looking at a case where the higher the R^2 value the better the model fits the data (minitab.com, 2013).

One of the limitations of R^2 values are they cannot determine whether the data plots are biased from the way the trendline fits the model. We have already determined this through visual inspection in the operator's graphical results, where we looked at how many points are above and below the trendline. Therefore, at this stage we can use the R^2 value to try to separate operator eight and ten.

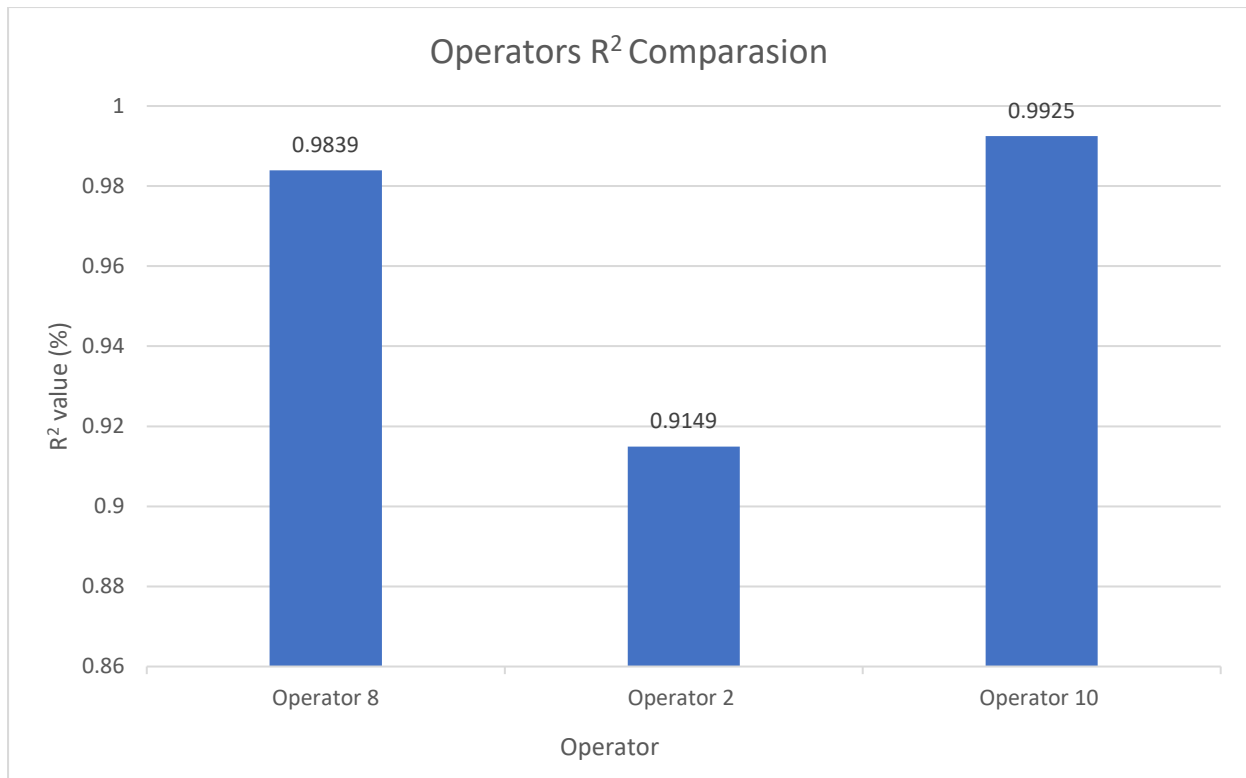


Figure 20 PoS R-squared comparison

The graph above indicates that operator ten has the highest R² value out of operators ten and eight, we have also plotted operator two in an effort to represent how a poor correlation of readings looks in terms of the R² value.

6.2.3 $mx + C$ Comparison

If we now look at the $mx + C$ values for the operators, we can try to determine who may have been out of calibration (from the C values) and who has the mx closest to 1 (which would follow the gradient of the true size).

The C value represents the y intercept on the graphs, this will help to determine whether an operator was out of calibration. For example, the true size C value will be 0. Therefore, if a graphs plots are intercepting the y axis consistently above or below 0 we could assume that there is a calibration issue, as true calibration will be defined as $c = 0$.

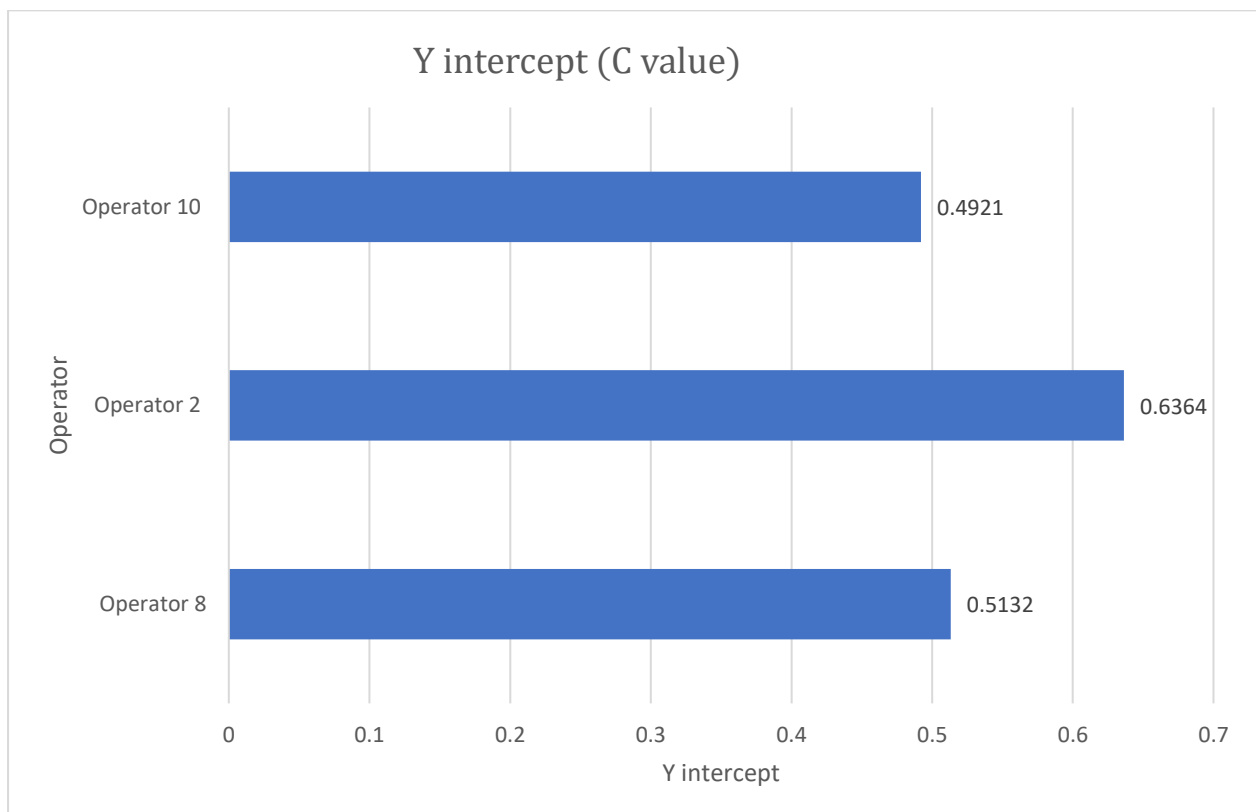


Figure 21 PoS y intercept comparison

Again, operator two has been displayed to show the worst-case scenario in relation to the most correlated data from operators eight and ten. We can see that on the whole operator ten has a Y intercept closer to zero (true size), than operator eight has recorded. This would indicate operator ten is using equipment with a higher degree of calibration in relation to the true size, and from the R^2 value has recorded a more consistent set of data.

6.2.4 Combining PoS Data

From the analysis thus far, we look to be drawing on the conclusion that operator ten has recorded the best set of results from the ultrasonic testing. A final area which we will investigate is the slope of the trendline (mx). We know that the true size slope is equal to 1, therefore we can assume that the graph with the closest value to one would be representing the true size with a higher degree of accuracy. We also know that when looking at the R² values we were looking for values as close to 1 (100%) as possible, because this indicated a close correlation of data around the trendline.

Therefore, for the two operators we have two values which we are comparing that we consider having an optimum number of 1. We can look at combining the R² value and the mx of each operator, finding the average of this figure and look to see who out of operator eight and ten has scored closer to the target number of 1.

$$\frac{R^2 + Mx}{2}$$

Operator eight;

$$\frac{0.9839 + 0.9945}{2} = 0.9892$$

Operator ten;

$$\frac{0.9925 + 0.8968}{2} = 0.94465$$

We can represent these results as a percentage on the next page;

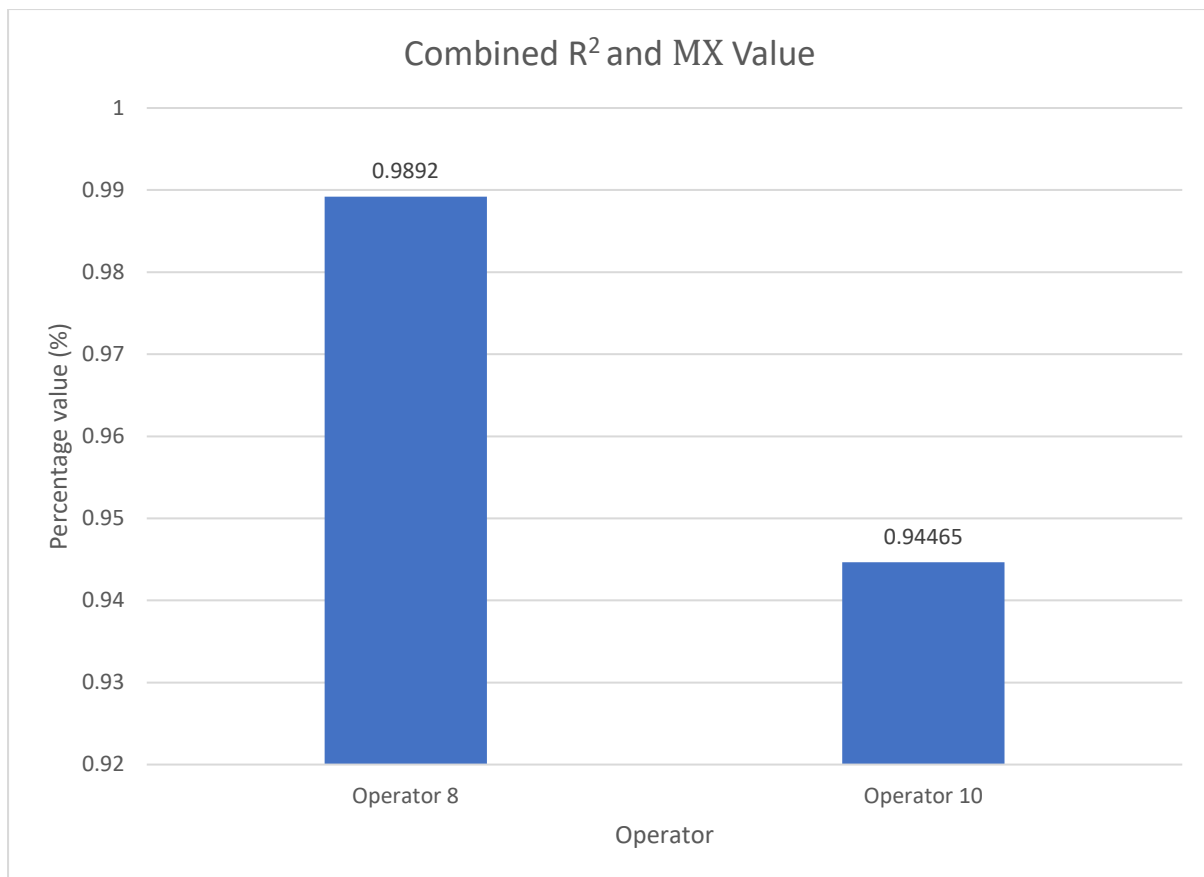


Figure 22 PoS Combining values Comparison

From this calculation it can be determined that operator eight has recorded the most accurate set of data. Although values were very close, it would indicate that operator eight can be separated from operator ten from the calculation and graph above, we could see that operator ten had slightly better calibration as there was a y intercept value recorded closer to zero than operator eight. But this was only a difference of 0.0211 in the y intercept, which makes it very difficult to draw on who has recorded the most accurate data from this alone. Once we determined the average sum of the R² and C values it separated operator eight from ten, and indicated that the data operator eight recorded contained a higher degree of accuracy.

6.2.5 PoS Operators Overview

We will now try to give a general overview of operator performance to try to determine in what significance some operators were away from the True Size readings.

Mean & Standard Deviation

To short list the operators into focused groups of high performers, the mean and standard deviations of the operator errors were calculated separately, combined together, then taken away from the True Size combination to see how far away the results were from the true size.

Table 4 Operator overview mean and standard deviations

	Mean	Standard Deviation	Combined Results	Combined – True size Combined
True Size	1.720	1.999	3.719	
Operator 1	1.933	1.998	3.805	0.086
Operator 2	1.929	1.570	3.499	-0.220
Operator 3	1.793	1.761	3.554	-0.165
Operator 4	2.110	1.700	3.810	0.091
Operator 5	2.029	1.895	3.924	0.205
Operator 6	2.414	2.044	4.458	0.739
Operator 7	1.733	1.698	3.432	-0.287
Operator 8	2.224	2.004	4.228	0.509
Operator 9	2.462	1.780	4.242	0.523
Operator 10	2.035	1.799	3.834	0.115

If we graphically plot this Operator Error from each operator in a bar graph we compare who's scores lie furthest from the Zero Position (considered as the True Size).

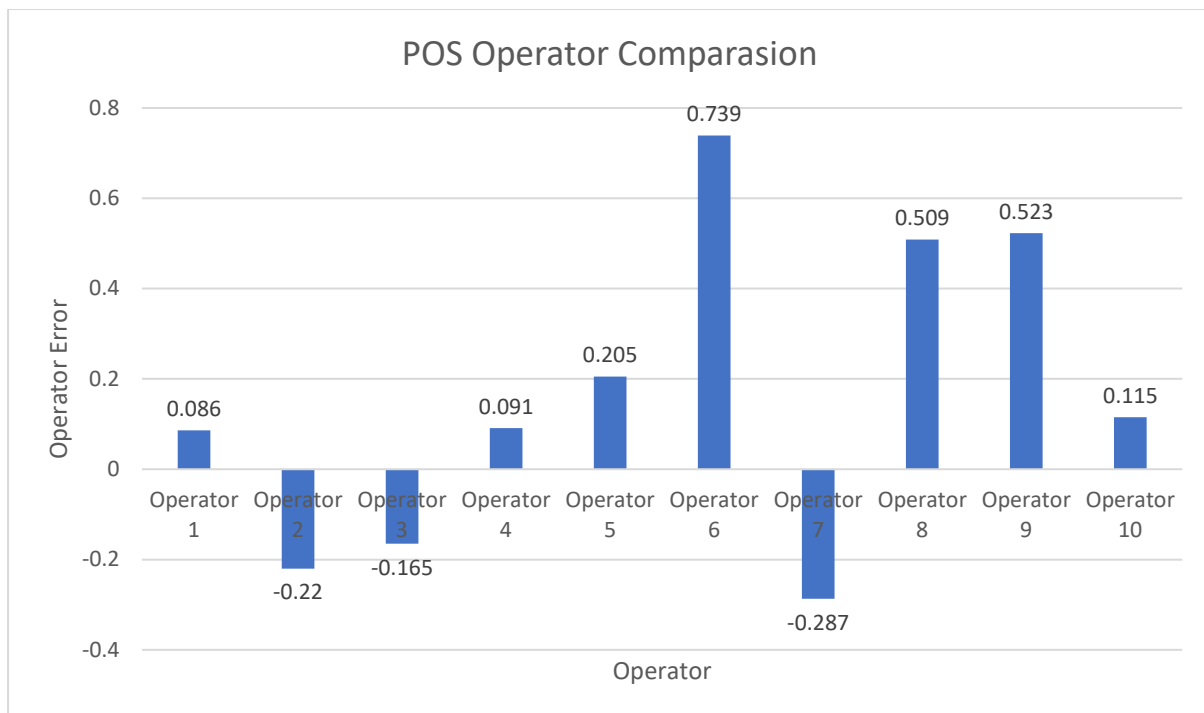


Figure 23 PoS operators mean & standard deviations

From the graph above, we can see that it suggests (if we are only taking into account the mean and standard deviation of recorded results) that operators one, four and ten have recorded the closest sets of results to the true size, with operators six, eight and nine being farthest from the True Size.

This should not be used to identify which operators recorded the most accurate results, this was determined in the sections previous, but is useful to identify the operator with the sizing estimates closest to the true size data set, as visually we previously analysed it that operator eight and ten were the two most successful. This information does not show in the above graph.

6.2.5.1 Operator Calibration

We previously mentioned calibration issues can be highlighted in the y intercept value. If we now compare these values of y intercept for all ten operators, we can look who is likely to be furthest from True Size calibration (which we deem to be zero).

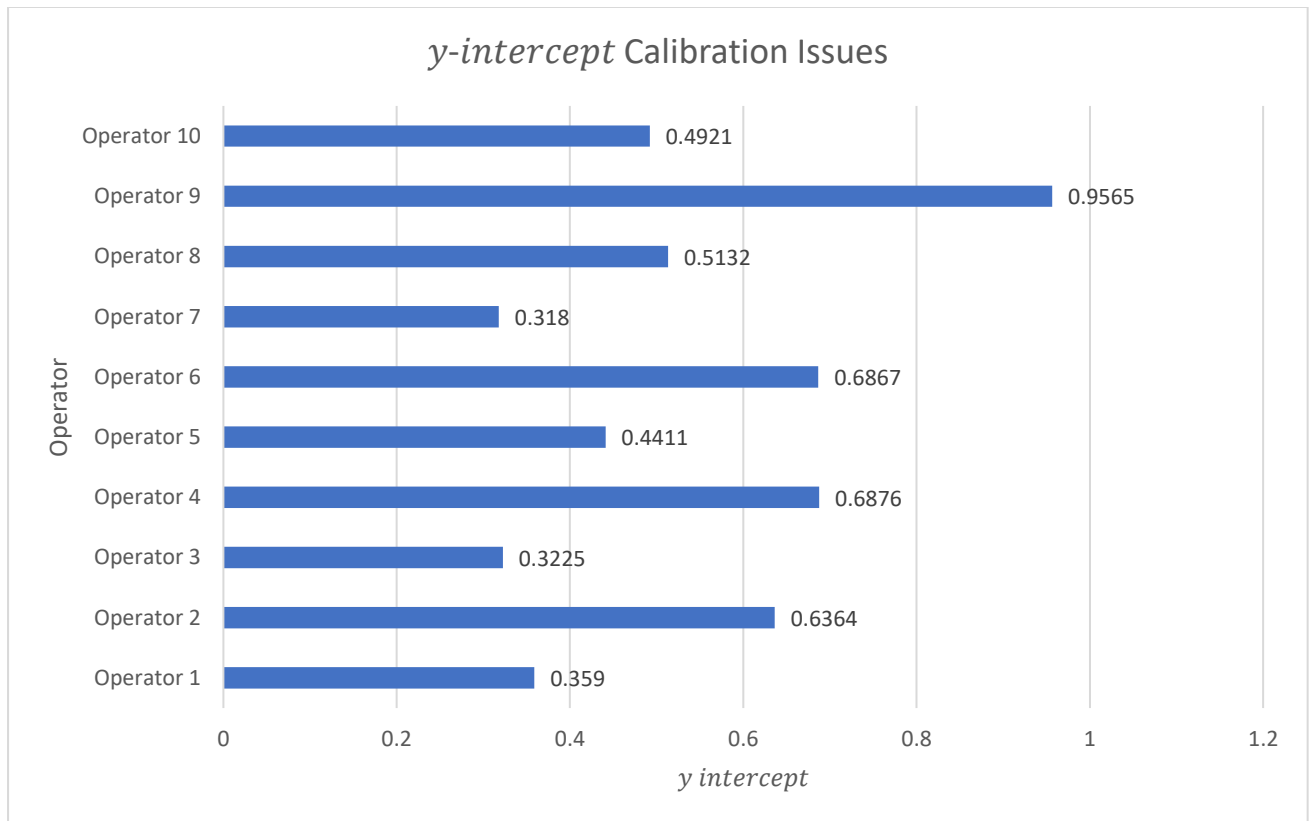


Figure 24 PoS y-intercept calibration issues

Here we can see that operators one, three and seven are running close to ideal calibration (0), therefore for them it would be hard to argue a case of equipment calibration as this small discrepancy could just be human error. But when we look at operators six, four and two we could start to argue for a calibration issue for the group as a whole as the results of these three differ from the rest of the group. An operator where we could make the statement with a good degree of confidence that there is a calibration issue is operator nine, the results displayed here do not follow a trend of any other operator, and the trendline indicates that everything is almost 1mm oversize consistently.

7. Discussion

By analysing operator recordings with statistical methods, it was proved that we could define which operators had recorded the most accurate data, when we can expect certain degrees of confidence from operators and how we can highlight possible calibration issues in equipment.

In the PoD section we could see how when the operators were recording consistent data as the defects got larger that a smooth curve was forming, with the most competent operators have faster rates of changes in their PoD curves. When we were looking at small sized defects (1mm) and under the operators struggled to record a hit, and displayed inconsistency until the defects started to become larger. We found that operator seven recorded the most accurate set of data, through research it was found that the operators were more likely to spot defects with the designated equipment and produce accurate PoD curves if they followed the set procedure diligently, although operator eight recorded similar results and even managed to find a smaller defect size than operator seven, the assumption could be made that operator seven overall followed the set procedure to finer detail.

The poor results displayed by operator four could be put down to operator inexperience or machine calibration, but as the lack of necessary hits in operator four's findings were graphically apparent we would have to assume it was down to either inexperience or machine calibration.

In section six we analysed the probability of sizing a defect for ten operators, by creating confidence limits graphically on the data we could immediately make the decision on which operators to analyse. The operators with the narrow confidence bands had this due to a small standard deviation, therefore the operators with narrow confidence bands were immediate points of interest.

After narrowing it down to a decision between two operators who had recorded very similar data we needed more than graphical methods to split the two and state who is more accurate. After analysing R^2 values we needed to further combine this R^2 values with the mx in order to make the

decision. We decided that operator eight produced the most accurate readings, but for us to make this decision it took a greater deal of analysis than was needed in PoD.

After inspecting calibration using the y-intercept values there was a pattern that emerged. We could see straight away that operator nine must have been out of calibration as the data recorded gave a y-intercept almost three times greater than operator seven, and generally bigger than the whole group. But graphically we could see that operators seemed to form three groups from their similarity's in y-intercepts, we see that operators six, four and two have a similar intercept, operators ten and nine have a similar intercept, and operators seven, three and one (see figure 24). Further investigation may need to be carried out to see if these groups of operators followed the procedure in a certain way together within their groups.

8. Conclusion

From the statistical analysis and research, it was proved that if PoD curves are understood you will be able to make informed decisions about the process i.e. how small a defect on average you can detect. This information could be further investigated by comparing the operator experience levels to defect dimensions, then conclusions could be drawn upon such as how user friendly the equipment is and would it take an operator of high skill level to find a certain size defect. But it should be remembered that for this study we were only provided with defect length and width dimensions, therefore this made analysis simplified with less variables to consider.

Again, with the PoS we could not make a reference to operator skill level or experience as to why certain operators performed better, but we were able to draw some conclusions upon the calibration of the machines which could be a cause for further investigation as the operators seemed to form in groups during the calibration.

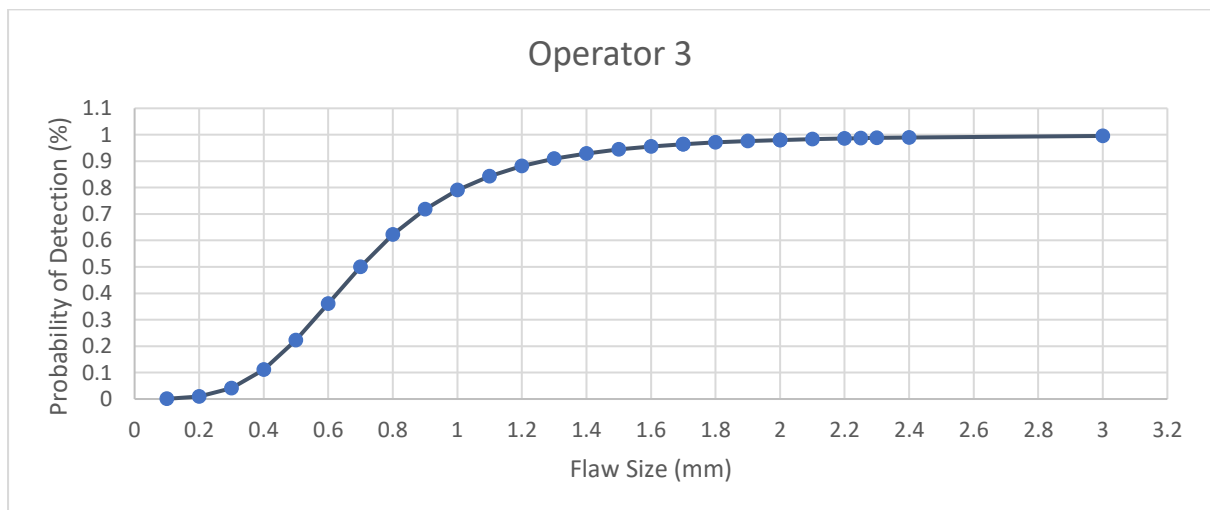
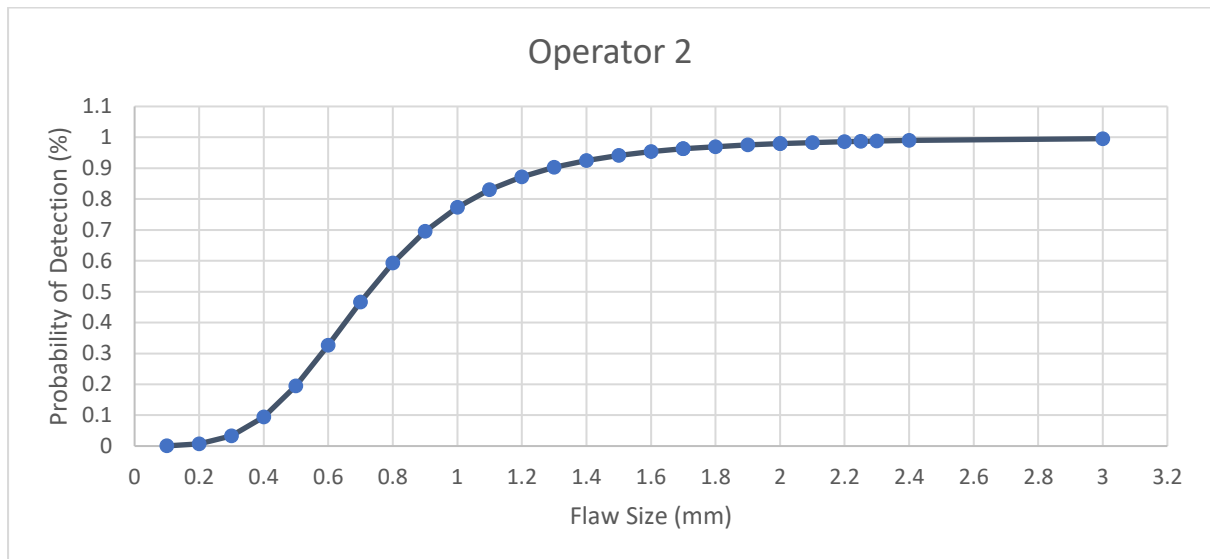
We were able to make clear statements on which operators recorded the best sets of data during the report (operator seven PoD, Operator eight PoS) with the methods displayed. If more variables were to be added to the data sets such as operator skill level and operator experience we would further be able to analyse the usability of equipment and complexity of the task.

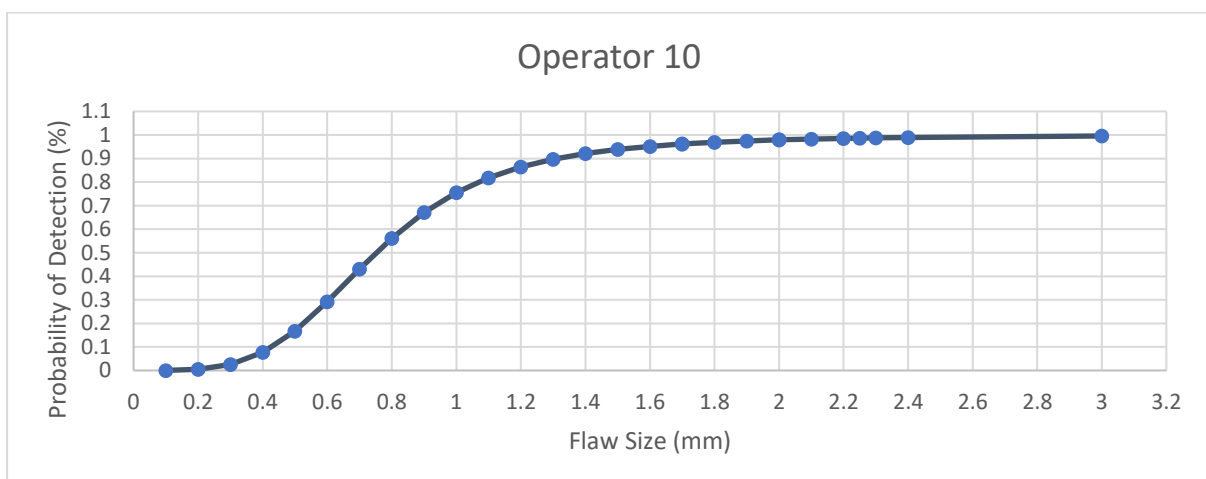
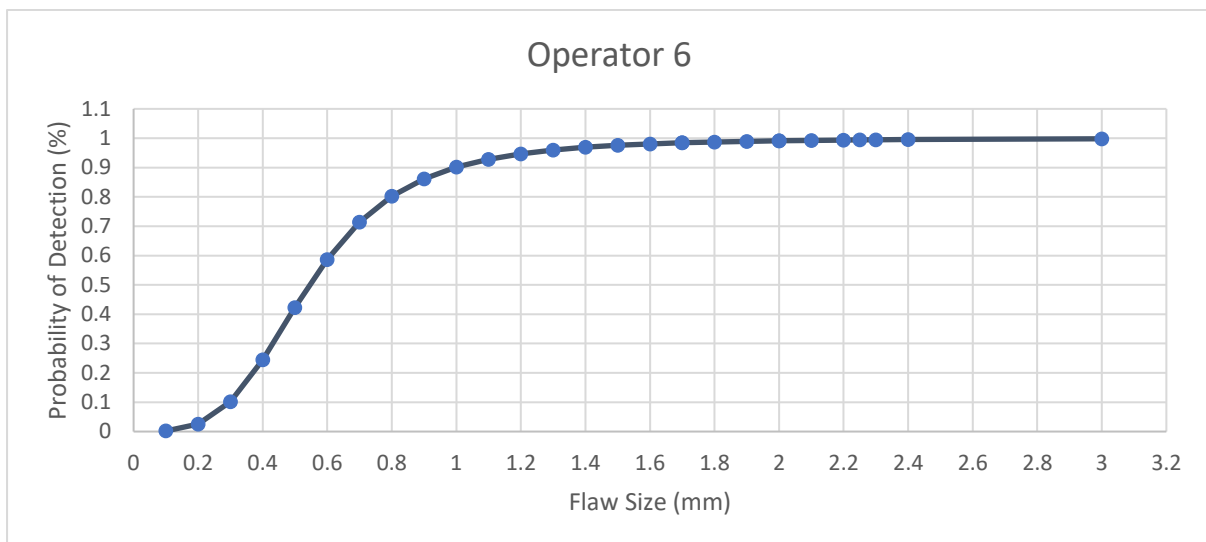
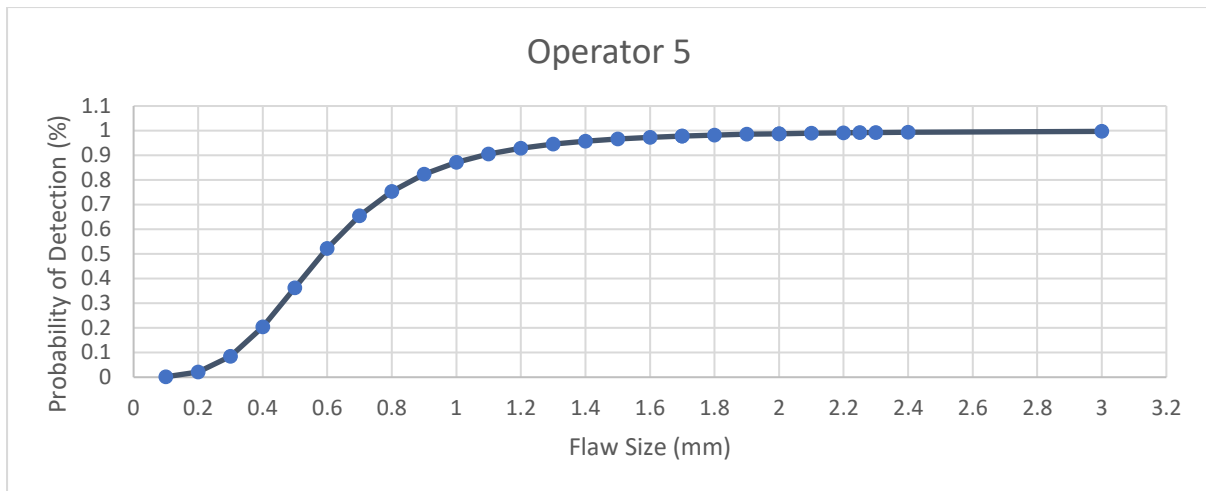
9. Bibliography & References

- B. Raj, . T. (2007). *Practical Non-Destructive Testing*. Moriston House, 75 Springfield Rd, Chelmsford CM2 6JB: Alpha Science International Ltd; 3rd Revised edition edition .
- Berens. (1997). NDE reliability data analysis. *non-destructive evaluation and quality control: qualitative non-destructive evaluation*.
- itl.nist.gov. (2013, 10 30). *itl.nist.gov*. Retrieved from itl.nist.gov:
<http://www.itl.nist.gov/div898/handbook/prc/section1/prc14.htm>
- Jacobi, G. A. (2006, January 2015 01). *Probability of Detection (PoD) curves HSE*. Retrieved from Probability of Detection (PoD) curves HSE: <http://www.hse.gov.uk/research/rrpdf/rr454.pdf>
- L.Phillips, J. (1973). *HOW TO THINK ABOUT STATISTICS* . New York, New York : Holt paperbacks .
- Lewis W H, S. W. (1978). Reliability of non-destructive inspection – Final Report. *Reliability of non-destructive inspection – Final Report*.
- minitab.com. (2013, May 01). *minitab.com*. Retrieved from minitab.com:
<http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
- people.duke.edu. (2001, 01 01). <https://people.duke.edu/~rnau/rsquared.htm>. Retrieved from <https://people.duke.edu/~rnau/rsquared.htm>:
<https://https://people.duke.edu/~rnau/rsquared.htm/~rnau/rsquared.htm>
- PISC-I. (1979). PISC-I. *Commission of the European Communities*, Volumes 1 to VI.
- R, D. W. (1996). Results of probability of detection trials. *Results of probability of detection trials*, 13-16.
- www.bindt.org. (2013, Januray 01). *www.bindt.org*. Retrieved from www.bindt.org:
<http://www.bindt.org/What-is-NDT/Index-of-acronyms/P/POD/>.
- www.ndt.net-POD. (2006, May 01). *www.ndt.net-POD*. Retrieved from www.ndt.net-POD:
<http://www.ndt.net/article/v11n05/ginzel2/ginzel2.htm>

10. Appendix

1.





2.

Defect	true size
A	0.25
B	0.27
C	0.28
D	0.4
E	0.4
F	0.5
G	0.52
H	0.55
I	0.7
J	0.98
K	1.12
L	1.15
M	1.1
N	1.3
O	1.6
P	1.8
Q	2.1
R	2.7
S	4.9
T	6
U	7.5
Mean	1.72
STD DEV	1.99897

3.

