U.S. Department
of Transportation

Federal Railroad
Administration

Office of Research,
Development and Technology
Washington, DC 20590

Analysis of Historical Non-Destructive Evaluation Probability of Detection Data for Railroad Tank Cars

DOT/FRA/ORD-21/14

Final Report | April 2021

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. Any opinions, findings and conclusions, or recommendations expressed in this material do not necessarily reflect the views or policies of the United States Government, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government. The United States Government assumes no liability for the content or use of the material contained in this document.

NOTICE

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.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY*
April 2021

2. REPORT TYPE
Technical Report

3. DATES COVERED *(From - To)*
1996 to 2016

4. TITLE AND SUBTITLE
Analysis of Historical Non-Destructive Evaluation Probability of Detection Data for Railroad Tank Cars

5a. CONTRACT NUMBER
DTFR53-11-D00008L

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Anish Poudel – 000-0002-5811-4284
Silvia Galván Núñez – 0000-0001-7943-8678
Brian Lindeman – 0000-0003-3903-266X

5d. PROJECT NUMBER

5e. TASK NUMBER
Task Order F000067

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Transportation Technology Center, Inc.
A wholly owned subsidiary of Association of American Railroads
55500 DOT Road
Pueblo, CO 81001-0130

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Railroad Administration
Office of Railroad Policy and Development
Office of Research, Development and Technology
Washington, DC 20590

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DOT/FRA/ORD-21/14

12. DISTRIBUTION/AVAILABILITY STATEMENT
This document is available to the public through the FRA website.

13. SUPPLEMENTARY NOTES
COR: Francisco González, III

14. ABSTRACT
From 1996 to 2016, Transportation Technology Center, Inc. (TTCI) analyzed historical probability of detection (POD) data for the fusion welded tank car butt welds (BW) and fillet welds (FW) inspection using Code of Federal Regulation (CFR)-approved nondestructive evaluation (NDE) methods. The research team analyzed a total of 197 POD datasets (i.e., for both the BW and FW) using 3 different statistical approaches: traditional statistics, logistic regression based maximum likelihood estimate approach, and the National Aeronautics and Space Administration (NASA) design of experiment POD (DOEPOD). The Federal Railroad Administration (FRA) sponsored this work and the railroad tank car industry participated by evaluating the capabilities of currently CFR-approved NDE methods on the railroad tank cars welds.

15. SUBJECT TERMS
Nondestructive evaluation/testing, NDE, nondestructive testing, NDT, visual testing, VT, liquid penetrant testing, PT, magnetic particle testing, MT, ultrasonic testing, UT, phased array ultrasonic testing, PAUT, probability of detection, POD, railroad tank car, butt weld, BW, fillet weld, FW, weld defect, hazardous materials, rolling stock

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Anish Poudel, Ph.D., Principal Investigator I (NDT)
Unclassified	Unclassified	Unclassified		54	19b. TELEPHONE NUMBER *(Include area code)* 719-584-0553

METRIC/ENGLISH CONVERSION FACTORS

ENGLISH TO METRIC

LENGTH (APPROXIMATE)

1 inch (in)	=	2.5 centimeters (cm)
1 foot (ft)	=	30 centimeters (cm)
1 yard (yd)	=	0.9 meter (m)
1 mile (mi)	=	1.6 kilometers (km)

AREA (APPROXIMATE)

1 square inch (sq in, in^2)	=	6.5 square centimeters (cm^2)
1 square foot (sq ft, ft^2)	=	0.09 square meter (m^2)
1 square yard (sq yd, yd^2)	=	0.8 square meter (m^2)
1 square mile (sq mi, mi^2)	=	2.6 square kilometers (km^2)
1 acre = 0.4 hectare (he)	=	4,000 square meters (m^2)

MASS - WEIGHT (APPROXIMATE)

1 ounce (oz)	=	28 grams (gm)
1 pound (lb)	=	0.45 kilogram (kg)
1 short ton = 2,000 pounds (lb)	=	0.9 tonne (t)

VOLUME (APPROXIMATE)

1 teaspoon (tsp)	=	5 milliliters (ml)
1 tablespoon (tbsp)	=	15 milliliters (ml)
1 fluid ounce (fl oz)	=	30 milliliters (ml)
1 cup (c)	=	0.24 liter (l)
1 pint (pt)	=	0.47 liter (l)
1 quart (qt)	=	0.96 liter (l)
1 gallon (gal)	=	3.8 liters (l)
1 cubic foot (cu ft, ft^3)	=	0.03 cubic meter (m^3)
1 cubic yard (cu yd, yd^3)	=	0.76 cubic meter (m^3)

TEMPERATURE (EXACT)

[(x-32)(5/9)] °F = y °C

METRIC TO ENGLISH

LENGTH (APPROXIMATE)

1 millimeter (mm)	=	0.04 inch (in)
1 centimeter (cm)	=	0.4 inch (in)
1 meter (m)	=	3.3 feet (ft)
1 meter (m)	=	1.1 yards (yd)
1 kilometer (km)	=	0.6 mile (mi)

AREA (APPROXIMATE)

1 square centimeter (cm^2)	=	0.16 square inch (sq in, in^2)
1 square meter (m^2)	=	1.2 square yards (sq yd, yd^2)
1 square kilometer (km^2)	=	0.4 square mile (sq mi, mi^2)
10,000 square meters (m^2)	=	1 hectare (ha) = 2.5 acres

MASS - WEIGHT (APPROXIMATE)

1 gram (gm)	=	0.036 ounce (oz)
1 kilogram (kg)	=	2.2 pounds (lb)
1 tonne (t)	=	1,000 kilograms (kg)
	=	1.1 short tons

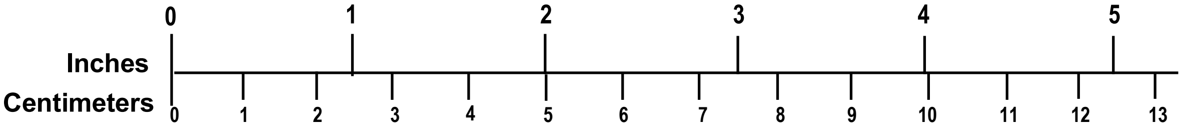
VOLUME (APPROXIMATE)

1 milliliter (ml)	=	0.03 fluid ounce (fl oz)
1 liter (l)	=	2.1 pints (pt)
1 liter (l)	=	1.06 quarts (qt)
1 liter (l)	=	0.26 gallon (gal)
1 cubic meter (m^3)	=	36 cubic feet (cu ft, ft^3)
1 cubic meter (m^3)	=	1.3 cubic yards (cu yd, yd^3)

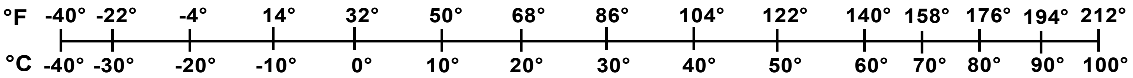
TEMPERATURE (EXACT)

[(9/5) y + 32] °C = x °F

QUICK INCH - CENTIMETER LENGTH CONVERSION



QUICK FAHRENHEIT - CELSIUS TEMPERATURE CONVERSIO



For more exact and or other conversion factors, see NIST Miscellaneous Publication 286, Units of Weights and Measures. Price $2.50 SD Catalog No. C13 10286

Updated 6/17/98

Acknowledgements

The authors would like to thank industry participants who supported this work. This research would not have been possible without the support and dedication of these individuals. Special thanks to Mr. Ward Rummel for his input and guidance during the early probability of detection (POD) development phase of this research. Mr. Rummel's expertise was invaluable to this research. Also, authors would like to thank Dr. Edward Generazio, National Aeronautics and Space Administration (NASA) nondestructive evaluation (NDE) program manager (retired) for his support on the design of experiment POD (DOEPOD) analysis.

Contents

Illustrations

Tables

Executive Summary

The Federal Railroad Administration (FRA) sponsored Transportation Technology Center, Inc. (TTCI) and accepted participation from the North American railroad tank car industry to conduct statistical assessments using data from 1996 to 2016. An evaluation took place to test the capabilities of Code of Federal Regulations' (CFR)-approved nondestructive evaluation (NDE) methods commonly used on the fusion welded railroad tank cars butt welds (BW) and fillet welds (FW). For this, TTCI fabricated and provided tank car panel test specimens with defects that were cut out from retired railroad tank cars. Defects in the BW and FW test panels were artificially created under cyclic loading conditions to directly imitate typical fatigue defects (cracks) found in tank car welds in revenue service.

Industry participation in this evaluation process consisted of 70 NDE operators (ASNT SNT-TC-1A certified Level I – Level III) from different companies that typically apply different NDE methods to inspect tank car FW and BW in revenue service, manufacturing, and repair environment. During the inspection process, operators used their own NDE inspection procedures, equipment, and inspection materials as they would in their own work in their normal work environment. The research team also briefed participants on the background, purpose, and the methodology of data collection and analysis. Finally, researchers gave each operator an incognito operator number during the testing, and the statistical data and graphs in this report reflect those numbers.

This report summarizes NDE results and the analysis of the probability of detection (POD) test results for all operators that participated in this study from 1996 to 2016. The POD curves included in this report provide a quantitative measure of the effectiveness of NDE methods, which provide an opportunity for a fleet owner to evaluate the need to use one method over another given the nature (criticality) of the area under observation, and the desired sensitivity. Results obtained also provide a baseline for each method so that changes to NDE variables become measurable by performing another study of the capabilities of the method and observing the resulting change.

1. Introduction

Transportation Technology Center, Inc. (TTCI), under the sponsorship of the Federal Railroad Administration (FRA) and with the support of the tank car industry, recently concluded studies to analyze the capabilities of current Code of Federal Regulations' (CFR)-approved nondestructive evaluation (NDE) methods and procedures to evaluate butt welds (BW) and fillet welds (FW) in railroad tank cars. For this, TTCI provided tank car panel test specimens with defects as well as provided supervision during NDE tests.

This report covers a comprehensive analysis of data collected during quantification of the CFR-approved NDE methods using different statistical metrics. In summary, the results indicate both the success and difficulties in applying some of the NDE methods to reliably detect and characterize fatigue cracks in the railroad tank car BW and FW. Data presented in this report also reflects operator and process variability in the application of the various inspection methods. The capability of any given NDE method or technique is specific to variables related to flaw characteristics such as size, orientation, and state of stress (compression or tension). The test object, inspection equipment, calibration, written procedure and related processes, acceptance criteria, human factors, and environmental conditions are all variables that affect NDE capability.

Results obtained from this research demonstrate that the CFR-approved NDE methods were not capable of achieving or approaching a 90-percent probability of detection (POD) with 95-percent confidence (90/95 POD) for fatigue cracks in the BW test panels, for the operators that participated in this research. Evaluation of the FW data showed mixed results, but only the magnetic particle testing (MT) method reached 90/95 POD. Also, excessive false calls were observed in both BW and FW inspection results. These results indicate the variability in NDE tests and calibration procedures, operator variance, and the influence of human factors in the application of the NDE inspection processes.

1.1 Background

The U.S. Department of Transportation (DOT) no longer considers the hydrostatic pressure test as an optimum way to qualify fusion welded tank cars for continued service. The main reason for this is due to the lack of ability of this test in identifying and characterizing fatigue damages in the tank cars resulting from in-service loadings, stress risers, and weld related defects (Garcia, G., 2002) (Garcia, G., Welander, L., Rummel, W. D., & Gonzalez, F., 2016) (Garcia, G., Rummel, W. D., & Gonzalez, F., 2016). Rulemaking issued by DOT revises the hazardous materials regulations (HMRs) to replace the hydrostatic pressure test with appropriate NDE methods to evaluate fusion welded tank cars. This rule change is contained in Title 49 CFR Section 180.509 (Code of Federal Regulations, 2012). Also, 49 CFR § 179.7 requires all tank car facilities to have a Quality Assurance Program (QAP), which is approved by the Association of American Railroads (AAR) and in compliance with AAR specifications for tank cars (Code of Federal Regulations, 2003) (Association of American Railroads, 2014). This rulemaking includes procedures for quantitatively evaluating inspection and test procedures, including an inspection of the accessibility of the area, and the sensitivity of the CFR-approved NDE methods. The changes in these regulations adopted the NDE methods for consistently, repetitively, and quantitatively detecting and characterizing internal defects and/or anomalies in the railroad tank car welds. The CFR currently authorizes the following NDE methods for tank car structural integrity inspections:

- Visual testing (VT)
- Liquid penetrant testing (PT)
- Magnetic particle testing (MT)
- Ultrasonic testing (UT)
- Radiographic Testing (RT)
- Acoustic Emissions (AE) (i.e., requires a special waiver from FRA)

This report summarizes the results of the 1996 to 2016 POD evaluations consisting of the CFR approved NDE methods and procedures used by industry personnel.

1.2 Objectives

The major objectives of this research are:

- To evaluate and quantify the capabilities of NDE methods authorized under 49 CFR § 180.509 for use in the qualification of railroad tank cars.
- To develop a quantitative POD approach to evaluate NDE techniques and increase the reliability of railroad tank car structural integrity inspections.
- To provide direction and insight into the current capabilities of the industry when using the allowed NDE methods.

1.3 Overall Approach

TTCI provided tank car panel test specimens and cut them out from retired railroad tank cars for the POD evaluations. Fatigue cracks were artificially simulated (under cyclic bending loading conditions) at the toe of the butt welds and at the longitudinal termination of the fillet welds in the cut-out sections of tank car panels. Cracks ranged from 0.15 inch to 3.50 inch for the butt weld panels and 0.10 inch to 4.50 inch for the fillet weld panels. A variety of cracks from smallest to largest sizes provided a range of inspection opportunities that were representative of cracked components from service.

Industry participation for the POD evaluations consisted of several NDE operators (ASNT SNT-TC-1A certified Level I – Level III) from different companies that usually apply different NDE methods to inspect tank car BW and FW in revenue service, manufacturing, and repair environments. During the inspection process, operators used their own NDE inspection procedures, equipment, and inspection materials as they would do in their normal work environment. Also, the operators briefed the participants on the background, purpose, and the methodology of data collection and analysis. Finally, each operator received an incognito operator number during the testing, and the statistical data and graphs presented in this report reflect those numbers.

The process implemented during tank car NDE POD evaluations also required each operator to inspect and size the EDM notches and fatigue cracks in the master gauge test panels before, at intervals during, and after completing the inspection of the larger tank car test panels. This was specifically done to aid the operators involved in POD evaluations to reinforce their familiarity with flaw responses from the test panels. Also, this served to ensure repeatability and reproducibility of the test process involved.

Researchers recorded the inspection results for the larger (blind) tank car test panels as hit or miss data for statistical analysis. For all BW inspections, NDE operators manually wrote the flaw size from start to end and location of the crack identified on a magnetic tape located from one end of the BW to the other end on each panel. A TTCI employee then measured and recorded the operator's response from the magnetic strips onto the data collection sheet. Subsequently, researchers entered all data results into the POD data template for further statistical analysis. Similarly, for all FW inspections, operators verbally identified the location of a crack and estimated its size, and a TTCI employee recorded the operator's response onto the data collection sheet. Subsequently, researchers entered all data results into the POD data template for further POD analysis.

1.4 Scope

TTCI conducted studies to evaluate a variety of CFR-approved NDE methods. The main goal of this study was to summarize prior work, understand the capabilities and limitation of currently approved CFR NDE methods and help the industry to achieve higher reliability of railroad tank car structural integrity inspections. FRA documented previous work in research reports (Garcia, G., 2002) (Garcia, G., Rummel, W. D., & Gonzalez, F., 2016) (Garcia, G., Welander, L., Rummel, W. D., & Gonzalez, F., 2016) (Archuleta, M., Poudel, A., Rummel, W. D., & Gonzalez, F., 2016). Previous work included manufacture and validation of physical tank car test specimens that are representative of components inspected by the industry as well as the results obtained from the prior POD studies. The use of these test specimens were to baseline industry detection capabilities. This report provides a comprehensive assessment of industry NDE inspector performance capabilities in detecting and characterizing fatigue cracks in the railroad tank car BW and FW using POD metrics.

1.5 Organization of the Report

This report presents the research findings in a progressive order. The next three sections present the results of the research methodology. Section 2 describes the research and test methodology implemented for this study. Section 3 provides insight into the consideration of background information on different statistical data analysis approaches for the analysis of historical NDE POD data obtained for the tank car BW and FW panels. Section 4 presents POD results obtained using different statistical data analysis approaches. Finally, Section 5 summarizes the work performed and provides recommendations for further work. The appendix lists all the data and detailed analysis results for the BW and FW panels.

2. Research Methodology

This section describes the research and test methodology implemented for this study.

2.1 Materials and Test Specimens

TTCI established a defect library containing sample artifacts, such as railroad tank cars and sections of railroad tank cars. Samples included tank cars donated by the tank car industry and manufactured artifacts developed at the FRA's Transportation Technology Center (TTC) in Pueblo, CO. Manufactured artifacts consisted of test panels used for POD study, along with master gauges developed for inspection sensitivity verification. The combination of specimens contains discontinuities developed in revenue service as well as manufactured flaws simulating locations and types of discontinuities expected in revenue service.

2.1.1 Tank Car Defect Library

TTCI developed realistic tank car panel test specimens that were cut from retired railroad tank cars (DOT 111A) for the POD evaluations. Tank car test panels are representative of ASTM A515 Grade 70 Steel material. Figure 1 and Figure 2 show a POD test setup for the tank car BW and FW test panels.

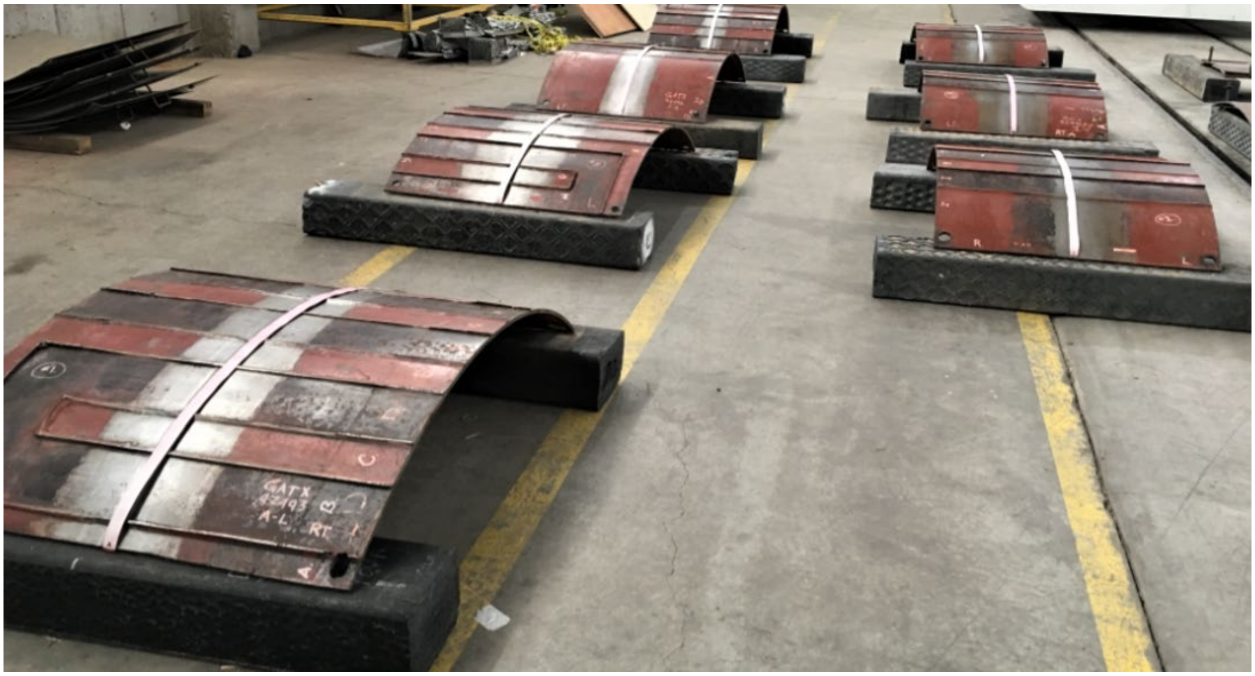


Figure 1. Tank Car BW POD Test Panels

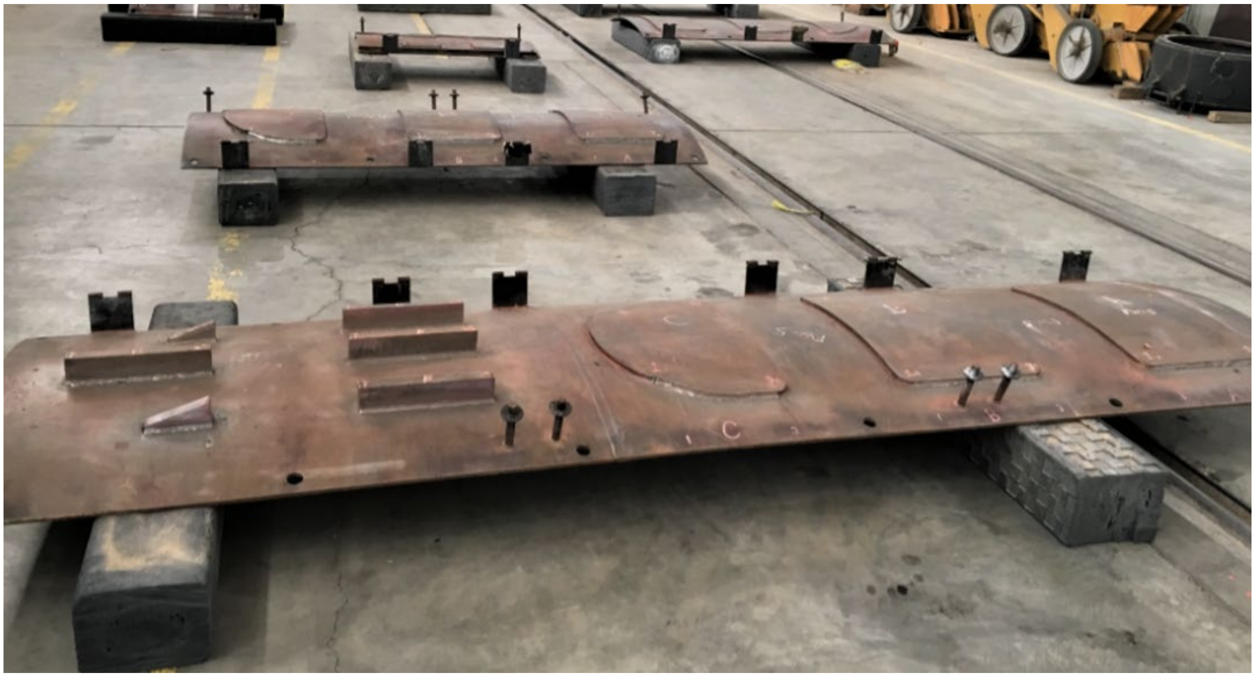
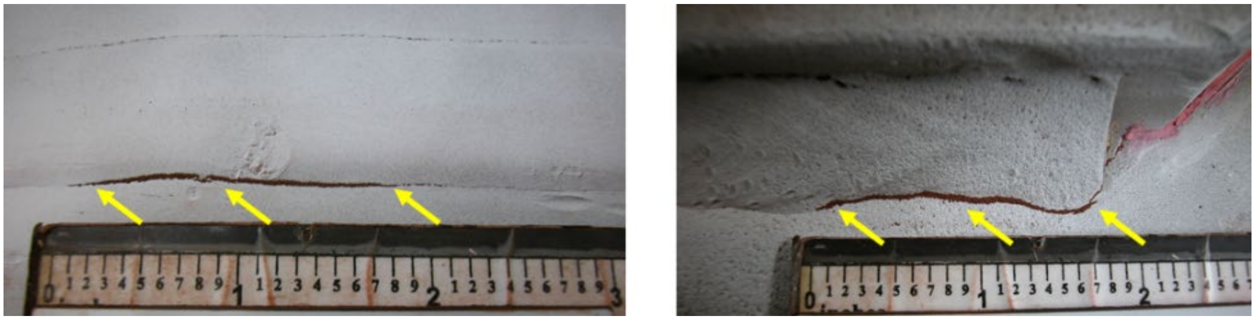


Figure 2. Tank Car FW POD Test Panels

For this test, researchers artificially initiated fatigue cracks (some were tightly spaced closed fatigue cracks) at the toe of the BW and at the longitudinal termination of the FW in the cutout sections of tank car panels as shown in Figure 3.



(a) **(b)**

Figure 3. Contrast MT Revealing Toe Cracking in Welds; (a) BW, (b) FW

Details on the tank car defect panel preparation can be found in the previous work conducted by Garcia et al. (Garcia, G., 2002) (Garcia, G., Rummel, W. D., & Gonzalez, F., 2016) (Garcia, G., Welander, L., Rummel, W. D., & Gonzalez, F., 2016). Development of the defect library provided the tank car industry with resources such as those established in the aerospace and nuclear industries. It offers the industry a facility to perform comprehensive, independent, and quantitative evaluations of existing approved NDE methods and test procedures, new and enhanced inspection, maintenance, repair techniques, and for operator training.

Figure 4 and Figure 5 show the distributions of the frequency of cracks (fatigue cracks) in the BW and FW tank car panels used in this study. The figures show that a right-skewed distribution describes the data, where smaller cracks had higher frequency than bigger cracks. It is called a right-skewed distribution because the location of the tail is on the right side. The selection of measure of central tendency (e.g., mean, median, and mode) for a skewed distribution might differ from a symmetric distribution. For a symmetrical distribution, the mean, median, and mode values lie on the same location. On the other hand, for a skewed distribution, the mode might not be a good representative because the location can be close to the left or right of the distribution. The mean value might be located at a place that is not the "center" of the distribution, although it might be close enough in some cases. Finally, the median value is a value that contains 50 percent of the data (50th percentile) and it is not susceptible to the skewness or outliers in the data (National Institute of Standards & Technology, 2012). In this research, the selection of median value was used as a measure of central tendency to the data.

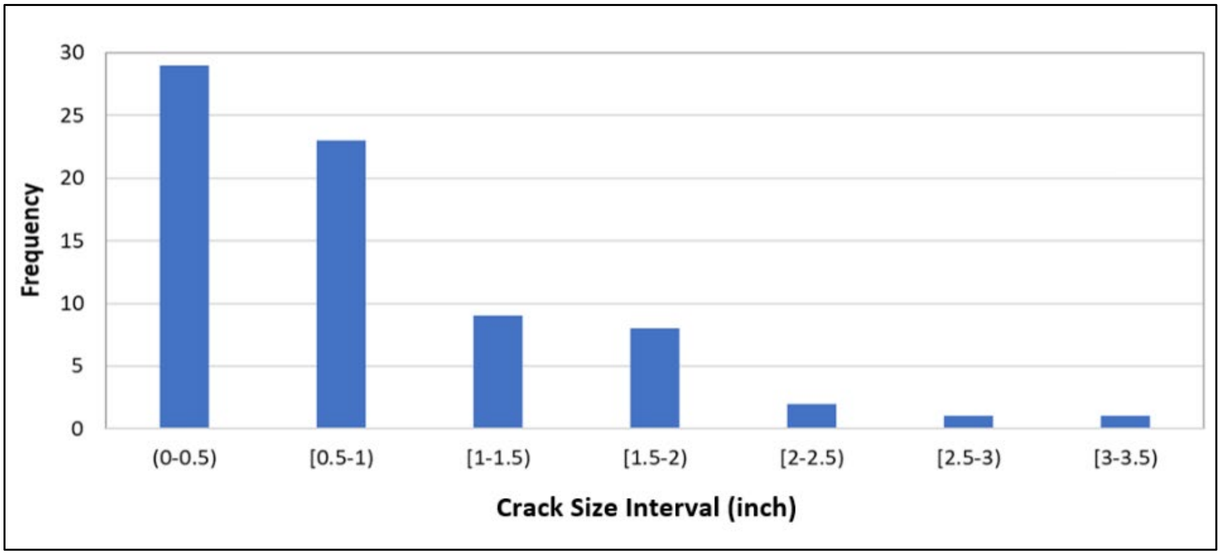


Figure 4. Distribution of Cracks for BW

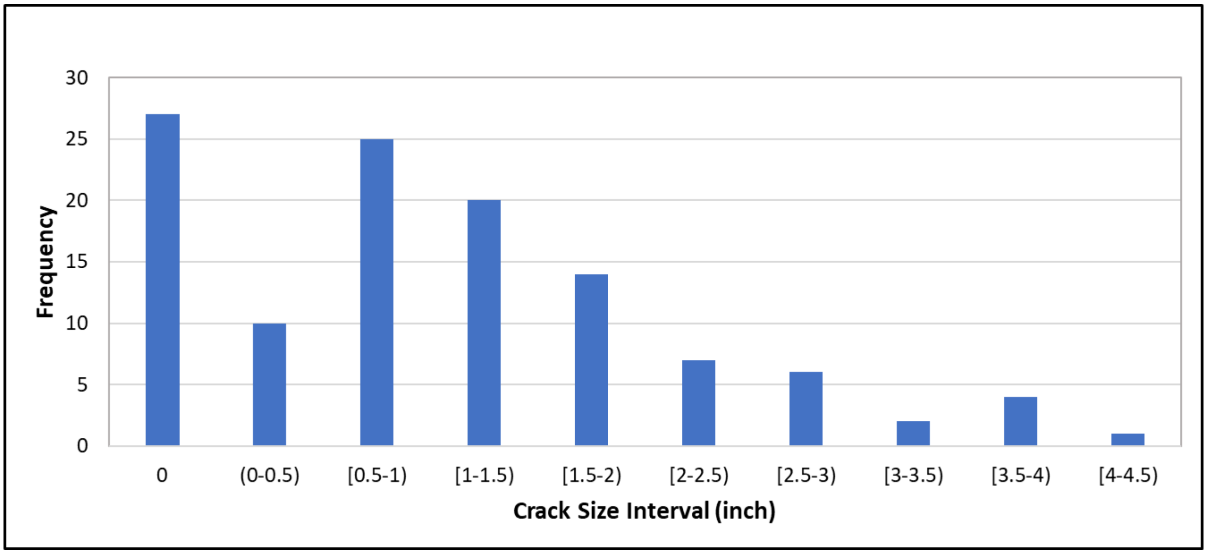


Figure 5. Distribution of Cracks for FW (dimension in inch)

2.1.2 Master Gauges

The creation of master gauges containing both notches, used electrical discharge machining (EDM) and fatigue cracks of varied sizes were manufactured by TTCI, as baselines for inspection sensitivity verification during the POD evaluations of industry NDE operators. The primary measures of reliability in NDE are repeatability (i.e., obtained through process control) and reproducibility (i.e., achieved through rigorous calibration). Unless reproducibility and repeatability are in control, NDE capabilities data (POD) is not in control and data is not representative of the inspection process. For NDE methods, such as PT and MT inspections, both the consistency of the inspection materials used and the sequence of application are critical to process repeatability. Similarly, for inspection methods, such as eddy current or ultrasound, which involves human pattern recognition and/or signal observation, there is a requirement for consistency in the threshold level used in detection (i.e., NDE process acceptance criteria).

Researchers developed master gauges from the test tank cars for use to perform a response comparison to calibration artifacts used in the field. The Transportation Technology Center (TTC) stores the master gauges to preserve and periodically revalidate response linearity of the calibration artifacts, as shown in Figure 6. For this study, each operator had a master gauge specimen every time before starting an assessment sequence to become familiar with the test specimen configuration and responses from the artificial fatigue cracks. The process implemented during tank car NDE POD evaluations required each operator to inspect and size the cracks and slots in the master gauge test panels before, at intervals during, and after completing the inspection of the larger tank car section panels.

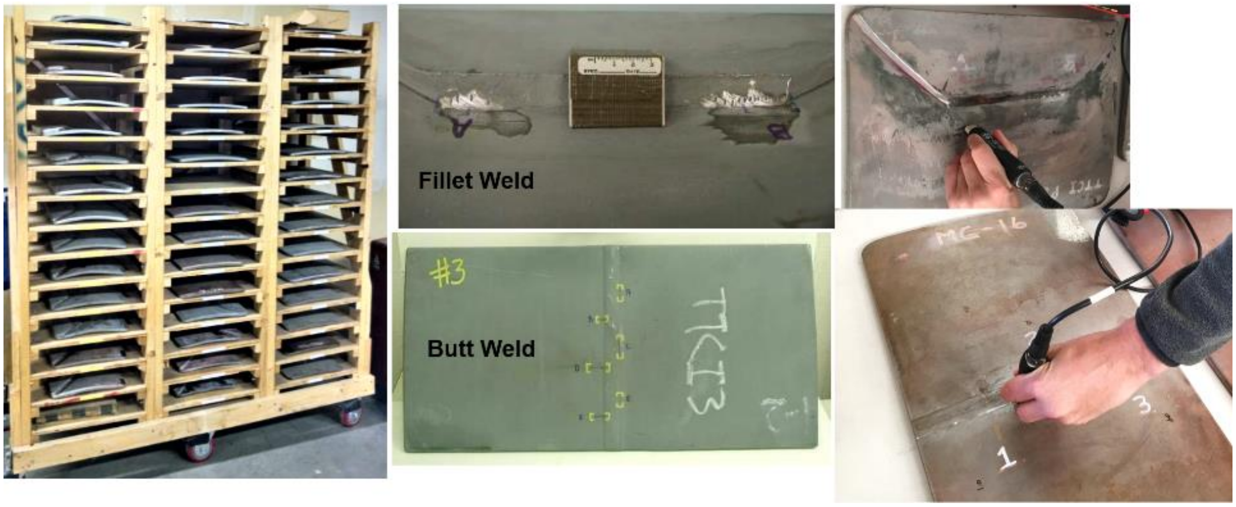


Figure 6. Tank Car Master Gauge Test Panels

2.2 NDE Methods

This project evaluated the performance/capability of the current NDE techniques for fatigue crack detection in railroad tank car fusion welds (BW and FW). Applicable methods were limited to VT, PT, MT, UT, and phased array ultrasonic testing (PAUT). A more in-depth description of the NDE methods used in this study can be found in the previous reports and outside literature (Garcia, G., 2002) (Garcia, G., Rummel, W. D., & Gonzalez, F., 2016) (Garcia, G., Welander, L., Rummel, W. D., & Gonzalez, F., 2016) (Archuleta, M., Poudel, A., Rummel, W. D., & Gonzalez, F., 2016).

2.2.1 Visual Testing

Performing VT took place with the unaided eye or with the use of some tools to enhance the detectability of discontinuities. The main advantage of the VT method is that it requires an operator to have limited training and equipment, whereas the main limitation of this method is the visual acuity of the observer or inspector.

2.2.2 Liquid Penetrant Testing

PT relies on capillary action principles where the liquid enters the surface cavities and later emerges as visual evidence of discontinuities such as defects within the panels. The main advantage of the PT method is that it is a rapid, simple method where large coverage is possible, whereas the main limitation of this method is subsurface discontinuities that are not exposed cannot be detected and characterized.

2.2.3 Magnetic Particle Testing

Usually, generating magnetic fields in test specimens takes place by direct or indirect magnetization processes. The underlying physics behind this technique is whenever there is a flaw in the test piece, it interrupts the flow of the magnetic lines of force, thus forming opposite magnetic poles. When the research team sprays fine magnetic particles onto the surface of the test specimen, the magnetic poles attract these particles, thus giving a visual representation of the indication. The advantage of the MT method is that it can detect surface and subsurface defects, whereas the limitations of this method are that it is only applicable to ferromagnetic materials and cannot be implemented if thick paint coating is present.

As with other NDE methods that use visual assessment to determine the integrity of the inspection area, MT can be enhanced by providing a greater contrast between the discontinuity and surrounding areas of the test article. Note that operators conducted the tests by both applying and by not applying a coating to the tank car specimen prior to inspection.

2.2.4 Ultrasonic Testing

In the UT approach, a thin layer of couplant is usually applied to the test object and the transducer scans over the part. This transducer sends out a pulse of energy and either the same or a second transducer listens for reflected energy (e.g., an echo). Reflections occur due to the presence of discontinuities and the surfaces of the test object. The main advantage of the UT method is that the depth of penetration for flaw detection or measurement is superior to other NDE methods. This makes it highly accurate in determining flaw location and estimating size and shape, whereas the main limitations of this method are that the surface must be accessible to transmit ultrasound and limitation to flaw detection capabilities due to fixed angled approach. In conventional UT, inspection parameters such as focal point and angle of incidence are mechanically fixed. The focal point in the material is the depth the inspection is performed, and the angle of incidence is the angle at which the ultrasonic signal is emitted into the material.

2.2.5 *Phased Array Ultrasonic Testing*

PAUT is an advanced ultrasonic NDE method that uses multiple elements (transducers) in a single probe housing with the capability to send an array of sound, in a wide range of angles, through the tested material. The main advantage of the PAUT method is that it uses multiple elements within a single transducer assembly to steer, focus, and scan beams which reduces inspection times and improves productivity, whereas the main limitations of this method is focusing the beam at a too shallow depth on the material, which means that deeper discontinuities may be missed.

2.3 NDE POD Data Collection

Industry participation for the POD evaluations consisted of 70 NDE operators (ASNT SNT-TC-1A certified Level I – Level III) from different companies that usually apply different NDE methods to inspect tank car BW and FW in revenue service, manufacturing, and repair environments. During the inspection process, operators could use their own NDE inspection procedures, equipment, and inspection materials as they would do in their normal work environment. Also, the research team briefed participants on the background, purpose, and the methodology of data collection and analysis. Finally, each operator received an incognito operator number during testing, and the statistical data and graphs in this report reflect those numbers. Table 1 and Table 2 shows the breakdown of all operators that participated for each NDE method and weld type. Tables 22 and 23 in Appendix I shows the breakdown of each individual operator who participated in multiple methods for both BW and FW panels.

Table 1. Number of Operator Participants for BW Panels Using Different NDE Methods

NDE Method	Number of Operators
VT	24
PT	25
MT with Contrast	19
MT without Contrast	11
UT	25
PAUT	3
Total	**107**

Table 2. Number of Operator Participants for FW Panels Using Different NDE Methods

NDE Method	Number of Operators
VT	26
PT	27
MT with Contrast	24
MT without Contrast	10
UT	3
PAUT	0
Total	**90**

The process implemented during tank car NDE POD evaluations also required each operator to inspect and size the EDM notches and fatigue cracks in the master gauge test panels before, at intervals during, and after completing the inspection of the larger tank car test panels. Specifically, this occurred to aid the operators involved in POD evaluations to reinforce their familiarity with flaw responses from the test panels. Also, this served to ensure repeatability and reproducibility of the test process involved.

Researchers recorded inspection results for the larger tank car test panels as hit or miss data for statistical analysis. The use of this data was used as an indicator of potential variation in the applied operator discrimination level during completion of the inspection sequences. When finding a large variation in discrimination and sizing, the false call number for that operator was usually high, and validity of the inspection sequence was therefore in question.

For all BW inspections, NDE operators manually wrote the flaw size from start to end and location of the crack identified on a magnetic tape located from one end of the BW to the other end on each panel. A TTCI employee then measured and recorded the operator's response from the magnetic strips onto the data collection sheet. Subsequently, the operator entered all data results into the POD data template for further statistical analysis. Similarly, for all FW inspections, operators verbally identified the location of a crack and estimated its size, and a TTCI employee recorded the operator's response onto the data collection sheet. Subsequently, all data results were entered into the POD data template for further POD analysis.

Researchers followed these guidelines to determine a hit, miss, or false call for the BW and FW panel nondestructive testing (NDT) evaluations:

1. Determining a hit: The location DOES contain an actual crack (any length) and the NDT operator finds a crack of any length within +/- 0.5-inch of the actual crack location, it would count as a hit.
2. Determining a miss: The location DOES contain an actual crack (any length) and the NDT operator DOES NOT find a crack of any length within +/- 0.5-inch of the actual crack location, it would count as a miss.
3. Determining a false call: If an NDT operator finds a crack of any length in a location that IS NOT within +/- 0.5-inch of the actual crack location, it would count as a false call.

Finally, related to human factors is the operator's ability to inspect an item within a given period, under a job quota and maintain production levels, thereby introducing an inherent need to inspect at a given rate. Consequently, the operator's ability to discriminate flaws at a standard inspection rate influences the POD curve. For example, if two operators evaluate a test sample, one operator may spend 15 minutes, while another operator may spend 30 minutes, depending on their comfort level for the decision-making process during flaw discrimination. Operator variability can also be seen for each operator depending on his/her status in the company and in the application of the various inspection methods and the effect of false calls on detection capability.

2.4 False Positive Indications

The NDE inspection process often challenges NDE operators to discern a flaw signal from the background response (noise) of the material that is inherent to the measurement. If the threshold discrimination is set too high, the operator will miss the flaw and the POD reduces. If the threshold discrimination is set too low, a false positive (i.e., noise interpreted as a signal) will result in instances where the signal and noise distributions overlap. The definition of a false call is this situation when an NDE operator identifies or records a flaw during an inspection that does not exist. False calls do not directly influence the POD curve (i.e., when based solely on a hit/miss approach). An operator could theoretically have a high POD and a correspondingly high false call rate. Optimal results should manifest a high POD with a low false call rate. Because false calls may lead to further inspections using additional NDE methods, fleet owners may experience costs associated with unnecessary maintenance, downtime, and repairs. Selection of the NDE method and technique should, therefore, be balanced between the POD results and the number of false calls.

3. Statistical Data Analysis Approach

Researchers considered three different statistical data analysis approaches for the analysis of historical NDE data obtained for the tank car BW and FW panels. The first approach focused on the traditional approach of calculating the probability of hits (POH) by obtaining the total number of hits in a given flaw size interval. This is to show the relationship between crack length ranges and the number of operators that obtained POH in the given crack length range. Note that this approach lacks accuracy because the resulting probability can be interpreted that each crack will have the same probability of being detected. In addition, for this POH to be true, it is important to account for an "infinite" amount of observations to obtain the "true" probability.

The second approach focused on calculating the POD values as a function of crack length. This approach uses the Logistic Regression (LOGIT) statistical model to estimate the parameters of a LOGIT model using the Maximum Likelihood Estimation (MLE) approach. The MLE is a frequentist method that calculates a posteriori probability of the parameters of a model given the observations, by maximizing the likelihood of observing the data given the parameters.

The third statistical data analysis approach is the National Aeronautics and Space Administration's (NASA) Design of Experiments for Probability of Detection (DOEPOD). DOEPOD uses a binomial distribution model for a set of flaws that are grouped into classes, where each class has a width. It also utilizes the concept of point estimate POH at a given flaw size and the lower confidence levels (LCL) of the observed estimated POH. It is the most conservative approach compared to the others mentioned earlier.

This section provides a quick background and history on the POD.

3.1 POD Background

A fatigue and fracture mechanics-based approach in the design, maintenance, and life extension of engineering systems provides quantification of confidence in the safety and structural integrity. Also, the emergence of a damage tolerance approach to determine inspection intervals for an engineered structure, such as railroad tank cars, requires the quantification of the detectable flaw size for the NDE methods used during inspection. The National Transportation Safety Board (NTSB) issued Safety Recommendations R-92-21 through R-92-24 suggesting a process of performing a reliable inspection of railroad tank cars based on a damage tolerance approach (National Transportation Safety Board, 1992). Damage tolerance design and maintenance requirements aim to improve the reliability and confidence level of tank car acceptance and maintenance.

A frequently used statistical metric to quantitatively measure the performance and capability of the NDE process/procedures is the POD. Researchers generated the POD graphs to relate the output of the NDE process/procedure to some characteristic of the test object, typically "cracks." This is done by subjecting a statistically significant number of flaws of varying size through an inspection procedure and plotting the detection/miss results as a function of flaw size (i.e., length, depth, depth-to-length ratio, depth-to-panel thickness ratio). However, many controllable and uncontrollable variables influence POD results. These include flaw characteristics (i.e., shape, size, and orientation), test object (i.e., material, thickness, and geometry), NDE methods/materials applied, NDE equipment, accept/reject criteria, NDE procedure/process/calibration, NDE personnel (i.e., education, training, experience), environmental condition, and human factors involved during the inspection process. In addition, repeated inspection of the same type and size of flaws also does not necessarily yield consistent results. There will typically be a spread in the detection results for the same flaw type and its size. These variations are inherent to any NDE process because of the variations in equipment setup, calibration, material properties, and flaw characteristics. Therefore, presenting NDE detection performance/capability is usually in a statistical term such as POD.

POD functions for quantifying the capabilities of NDE technique have been the subject of various investigations and have also experienced impressive advancement since its inception in late 1960s and early 1970s by NASA for its Space Shuttle Program (Pettit, D. E., & Hoeppner, D. W., 1972) (Rummel, W. D., Todd, P. H., Frecska, S. A., & Rathke, R. A., 1974) (Rummel, W. D., Rathke, R. A., Todd, P. H., & Mullert, S. J., 1975) (Rummel, W. D., Rathke, R. A., Todd, P. H., Tedrow, T. L, & Mullen, S. J, 1976), and followed by the Air Force aircraft programs (Lewis, W. H., Dodd, B. D., Sproat, W. H., & Hamilton, J. M., 1978) (Berens, A. P., & Hovey, P. W., 1983). The POD is now considered to be a standard method for demonstrating the capability of NDE processes and is widely accepted and integrated by many industries and agencies. Two standard approaches for analyzing the NDE test data and producing POD graphs have been proposed and both include (a) hit/miss data (binary response); and (b) a/â (quantitative signal response) (U.S. Department of Defense, 2004). In the hit/miss approach, recording the outcome of NDE results is a binary value, i.e., whether the flaw was detected (1) or not (0). Similarly, in the a/â approach, NDE signal response (i.e., â, 'a hat data') is recorded and is related to the flaw size (*a*). Next, hit/miss or a/â data are analyzed using different probabilistic statistical models to produce the POD(*a*) function. Some of the standard approaches include LOGIT, probit regression model (PROBIT), Bayesian, and Binomial Point Estimate Methods. Details on these probabilistic statistical models are well described in the literature (Rummel, W. D., Todd, P. H., Frecska, S. A., & Rathke, R. A., 1974) (Rummel, W. D., Rathke, R. A., Todd, P. H., & Mullert, S. J., 1975) (Rummel, W. D., Rathke, R. A., Todd, P. H., Tedrow, T. L, & Mullen, S. J, 1976) (Lewis, W. H., Dodd, B. D., Sproat, W. H., & Hamilton, J. M., 1978) (Berens, A. P., & Hovey, P. W., 1983) (U.S. Department of Defense, 2004) (Generazio, E. R., 2009) (Generazio, E. R., 2011) (Generazio, E. R., 2014).

The two parameter LOGIT model assumes that the POD is always increasing with the discontinuity size and is commonly expressed as:

$$POD(a) = \frac{exp^{\alpha + \beta \ln(a)}}{1 + exp^{\alpha + \beta \ln(a)}} \tag{1}$$

where, ɑ is the discontinuity size and α/β are the two parameters that are to be estimated using the MLE procedure. Although the function shown in Equation 1 describes a cumulative distribution (i.e., assumes a random Gaussian distribution); this function should not be confused with cumulative probability functions and the discontinuity size is not a random variable. The LOGIT model may also be described as:

$$POD\,(a) = F\left[\alpha + \beta\left(Log(a)\right)\right] \tag{2}$$

where, α/β are parameters to be fit to the data and F is an increasing function with respect to crack size, *a*.

A standout amongst the usually acknowledged metric of a sufficient NDE inspection process is that there should be 90 percent or greater probability of detection with 95 percent confidence for a given flaw size and greater (90/95 POD) (Generazio, E. R., 2009). The origin of implementing 90/95 POD as a metric for NDE inspection capability, derived from Mil-HDBK-5H where the 90/95 bound (T_{90} value) for acceptable B-basis material properties defined by U.S. Department of Defense (1998). The T_{90} is the value at which at least 90 percent of the population is expected to equal or exceed with 95 percent confidence. Figure 7 shows that the statistically computed value of T_{90} which represents a 95 percent LCL on the 10th percentile of the distribution; using a confidence limit assisted with calculating the value to provide a margin in the POD value.

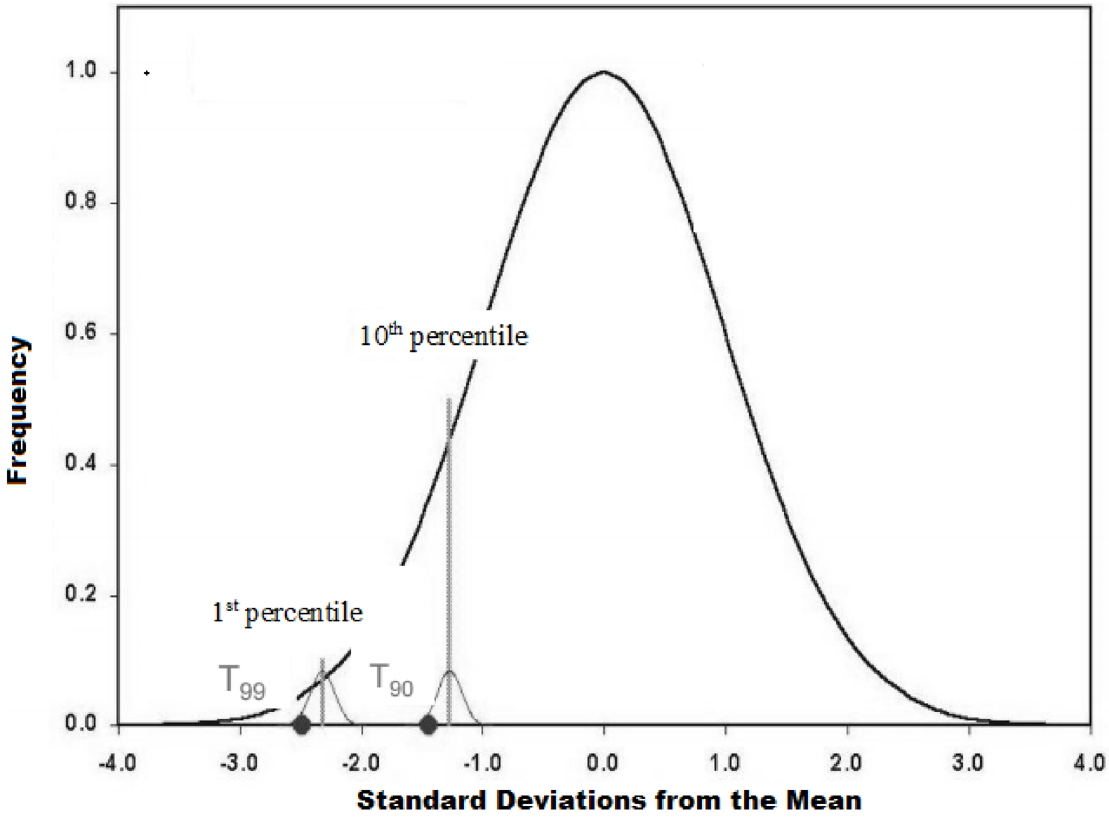


Figure 7. Normal Distribution Showing 1st and 10th Percentile Distribution for Computing T_{99} and T_{90} Values

Note: if the sample cannot be described by a normal or Weibull distribution, the T99 and T90 values must be computed by nonparametric (distribution free) means, which can only be done if there are at least 299 observations.

Maximum false call percentage (FCP) of 5 percent is allowed for use with multi-parameter MLE curve fits. The Advisory Group for Aerospace Research and Development Group (AGARD) suggested this 5 percent FCP for use with multi-parameter MLE curve fits (Advisory Group for Aerospace Research and Development Group, 1993). Note that the acceptance of FCP of 5 percent for use in multi-parameter MLE curve fits is still not clear whether this suggestion was for LOGIT, PROBIT, or for a/â curve fit approaches and whether this suggestion was for Wald or Likelihood ratio bounds. These are all important in deciding this acceptance level.

The first POD study conducted under a NASA program generated 118 Al 2219-T87 test panels containing 328 tightly closed fatigue cracks of varied size (Rummel, W. D., Todd, P. H., Frecska, S. A., & Rathke, R. A., 1974). The use of the binomial point estimate approach for hit/miss data was to generate POD graphs with confidence level as a function of crack size. The data was insufficient to plot the 95 percent confidence level (i.e., 60 observations), therefore, researchers computed the 90 percent confidence level (i.e., 29 observations in each data group). These data size requirements for POD assessment triggered several investigations in this field by various researchers seeking to develop alternative analysis procedures. Two notable procedures were then developed using smaller data sets and evolved as the baseline methods for use. These include NASA 29/29 procedure and Berens (LOGIT/PROBIT) model procedures (Rummel, W. D., April 16-20, 2010).

The initial NASA approach for generating the POD, also described in the tutorial handbook (Rummel, W. D., 1997), has been the foundation of the railroad tank car NDE POD work performed under the sponsorship of FRA. This method established many of the requirements in current specifications and identified as a possible goal for use in railroad tank car NDE inspections during the initial discussions of the HM-201 rulemaking.

3.2 DOEPOD

DOEPOD is a methodology based on the design of experiment (DOE) and is implemented via software to provide a detailed analysis of POD test data, guidance on establishing data distribution requirements, and resolving test issues (Generazio, E. R., 2009). It uses a binomial distribution model for a set of flaws grouped into classes, where each class has a width. It also utilizes the concept of point estimate POH at a given flaw size and the LCL of the observed estimated POH (Generazio, E. R., 2015). DOEPOD expands the prior NASA POD work based on binomial distribution by including the concept of LCL for establishing that there is 95 percent confidence that the POD is greater than 90 percent (90/95 POD). DOEPOD, moreover, fulfills the requirement for critical applications where validation of NDT systems, procedures, and operators are required even when a predicted POD curve is estimated (Generazio, E. R., 2015). DOEPOD does not assume random Gaussian distribution about the value to predict POD value like multiple-parameter curve fitting or model optimization approaches. The detailed description on the DOEPOD methodology, concepts, confidence bound, and false call rate analysis are well described in the literature (Generazio, E. R., 2009) (Generazio, E. R., 2011) (Generazio, E. R., 2014) (Generazio, E. R., 2015).

During operation, DOEPOD statistically analyzes the NDT inspection data, identifies different cases for the results obtained, and provides direction on what to do next depending on the case, including how to modify the DOE to continue to efficiently validate the inspection system. For example, if 90/95 POD is reached at a given flaw size, then DOEPOD will direct the operator to identify locations that need additional validation for other flaw sizes. If 90/95 POD is not reached, then DOEPOD will use best lower confidence value to identify where options are available to reach 90/95 POD. DOEPOD classifies the POD result into one of seven different cases, such as CASES 1, 2, 4, 5, 6, 7,[1] and survey datasets. Once the case is determined,

[1] CASE 3 (i.e., multiple discontinuity sizes where 90/95 POD is observed for a fixed class width) and CASE 0 (i.e., all hits) are included in CASES 1 and 2 in DOEPOD.

DOEPOD provides recommendations which, if successfully pursued, will help for the full system validation. Table 3 lists the DOEPOD analysis summary and recommendations for all cases.

Finally, DOEPOD yields a warning when the upper confidence bound of the FCP exceeds 0.03448. The observed 90/95 POD results, when the upper confidence bound of the FCP exceeds 0.03448, is not considered valid. The 3.448 percent is the quantitative upper Clopper-Pearson 95 percent bound at which the probability of false call (FCP) may produce a "Lucky Hit" that is added to the Number of Hits, resulting in an erroneous higher estimate of the POD.

Table 3. Summary of all Cases and Actions in DOEPOD (Generazio, E. R., 2011)

CASES	90/95 POD at X_{POD} reached?	Does X_{POH} exist?	Is POH = 1 everywhere > X_{Best_LCL}?	Is $X_{POH} \leq X_L/3$?	Large Flaw Validation Complete?	DOEPOD Analysis Summary and Recommendations
CASE 1	YES	YES	YES	N/A	YES	90/95 POD at X_{POD} has been reached. Actions: Address any false call warnings.
CASE 1+	YES	YES	NO	N/A	YES	90/95 POD at X_{POD} has been reached. Actions: Misses above X_{POD} need to be explained and resolved. Address any false call warnings.
CASE 1#	YES	YES	YES	N/A	NO	90/95 POD at X_{POD} has been reached. Actions: Further validation at flaw sizes greater than X_{POD} is required. Add large flaws. Address any false call warnings.
CASE 1*	YES	YES	NO	N/A	NO	90/95 POD at X_{POD} has been reached. Actions: Further validation at flaw sizes greater than X_{POD} is required. Add large flaws. Misses above X_{POD} need to be explained and resolved. Address any false call warnings.
CASE 2	YES	YES	NO	N/A	N/A	90/95 POD at X_{POD} has been reached. However, there are excessive number of Misses above X_{POD}. Actions: Add validation at identified flaw sizes is required. Add flaw per instructions.
CASE 4	NO	YES	YES	N/A	N/A	90/95 POD at X_{POD} has not been reached. Actions: Increase number of flaws at X_{POH} = 1 or X_{Best_LCL}
CASE 5	NO	YES	NO	YES	N/A	90/95 POD at X_{POD} has not been reached and there are misses above X_{Best_LCL}. Actions: Increase number of flaws at X_{POH} = 1
CASE 6	NO	YES	NO	NO	N/A	90/95 POD at X_{POD} has not been reached. The POH is fluctuating above X_{Best_LCL} and X_{POH} is greater than $X_L/3$. The inspection system is unstable for the flaw size range analyzed. Actions: Increase the flaw size range by a factor of two.
CASE 7	NO	NO	N/A	N/A	N/A	90/95 POD at X_{POD} has not been reached. The inspection system is unstable for the flaw size range analyzed. Actions: The inspection system may not be appropriate or increase the flaw size range by a factor of two.
Survey Cases	NO	YES	N/A	N/A	N/A	The optimized class width exceeds 1/3 X_L and X_{POD} has not been reached. The class width optimization has determined that there is a class width for which the smallest X_{POH} = 1 class length is identified. Actions: Add flaws at Survey/Optimum X_{POH}

4. POD Results and Analysis

This section presents results obtained using traditional statistics, LOGIT/MLE, and NASA DOEPOD approaches.

4.1 Traditional Statistics

The traditional statistics method consists of calculating the POH and probability of misses (POM) by obtaining the total number of hits or misses in a given flaw size interval and dividing it by the total number of observations in that interval. For this study, the flaw interval was set to 0.5 inches.

One of the major drawbacks observed of utilizing this method for this study is the sample size. As mentioned above, this method consists of calculating the observed frequency of hits, misses, and false calls by dividing the number of hits by the total number of observations. The resulting probability can be interpreted that each crack will have the same probability of being detected; meaning that for example, the probability of detecting a 4-inch crack is the same as detecting a 0.01-inch crack, which it might lack of applicability in the practical sense. In addition, for this POD to be true, it is important to account for an "infinite" or large enough number of observations to obtain the "true" probability. For this case, researchers calculated the POH by sub setting the data into 0.5-inch crack size intervals and the total number of data points did not exceed 30 observations.

4.1.1 FW

This section summarizes the median POH for each crack interval for FW for all operators and for each NDE method applied. To conduct the comparison of the different methods, an understanding that there are differences in the number of operators that participated in each method and the number of cracks observed in each crack length range must also be considered. Table 4 shows the summary table for the median POH as well as the total number of observations in each crack length range, and Figure 8 shows the median POH for each NDE method.

Table 4. Median POH Summary for the FW

Crack Length Range	Number of Observations	Median Probability of Hits - FW				
		VT	PT	MT-Contrast	MT-No Contrast	UT
0	27					
(0-0.5)	10	60%	60%	80%	100%	60%
[0.5-1)	25	44%	68%	88%	92%	84%
[1-1.5)	20	60%	70%	90%	98%	90%
[1.5-2)	14	61%	86%	100%	100%	93%
[2-2.5)	7	57%	86%	100%	100%	86%
[2.5-3)	6	50%	83%	100%	100%	100%
[3-3.5)	2	100%	100%	100%	100%	100%
[3.5-4)	4	75%	75%	100%	100%	100%
[4-4.5)	1	100%	100%	100%	100%	100%

Based on the results presented, it can be observed that VT method has the lowest POH compared to the other methods for crack sizes less than 3-inch as well as in the crack length range 3.5- to 4-inch. This is important to point out as it indicates that some of the operators using this method did not identify some cracks in this range. PT was the second method that provided equal or lower POH compared to the other methods, excluding crack length range 3.5 to 4 inches. In addition, MT without the Contrast method consistently showed equal or higher POH for all crack length ranges compared to the other methods. Appendix A presents a more detailed description of the number cracks, hits, misses, and false calls for each crack size interval and inspection method for FW. The median number of hits for each crack distribution displayed the POHs on top. Appendix A also presents another way to visualize the same traditional statistics data using box plots for FW.

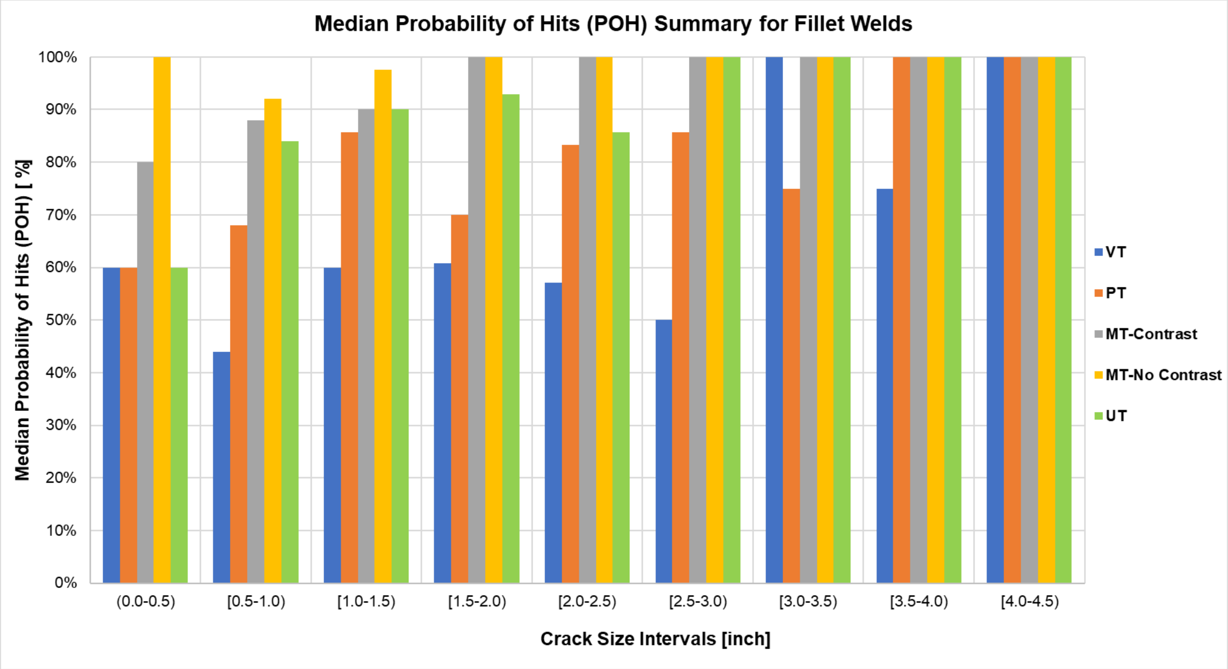


Figure 8. Median POH Summary for the FW

4.1.2 BW

This section presents the summary of the median POH for each crack interval for BW for all operators and for each NDE method applied. As mentioned in the previous section for FW, the comparison of the different methods also needs to be conducted with the understanding that there are differences in the number of operators that participated in each method and the number of cracks observed in each crack length range. Table 5 summarizes the POH as well as the total number of observations in each crack length range, and Figure 9 shows the median POH for each method. Per Table 5 and Figure 9, the VT method has the lowest POH compared to the other methods for crack sizes between 0.5 and 2.0 inches. The UT method reported the smallest POH for the crack length interval 0 to 0.5 inches. All the methods reported a POH of 100 percent in the crack length interval 3 to 3.5 inches and VT and PT methods had a 0 percent POH for crack length interval (2.5 to 3 inches). Appendix A presents a more detailed description of the number of cracks, hits, misses, and false calls for each crack size interval and inspection method for BW. The median number of hits for each crack distribution displays the POHs on top. Appendix A also presents another way to visualize the same traditional statistics data using box plots for BW.

Table 5. Median POH Summary for the BW

Crack Length Range	Number of Observations	Median Probability of Hits - BW					
		VT	PT	MT-Contrast	MT-No Contrast	UT	PAUT
(0-0.5)	29	38%	38%	55%	55%	34%	52%
[0.5-1)	23	35%	52%	65%	61%	61%	83%
[1-1.5)	9	56%	67%	78%	78%	67%	78%
[1.5-2)	8	75%	88%	88%	100%	88%	100%
[2-2.5)	2	100%	100%	100%	100%	100%	100%
[2.5-3)	1	0%	0%	100%	100%	100%	100%
[3-3.5)	1	100%	100%	100%	100%	100%	100%

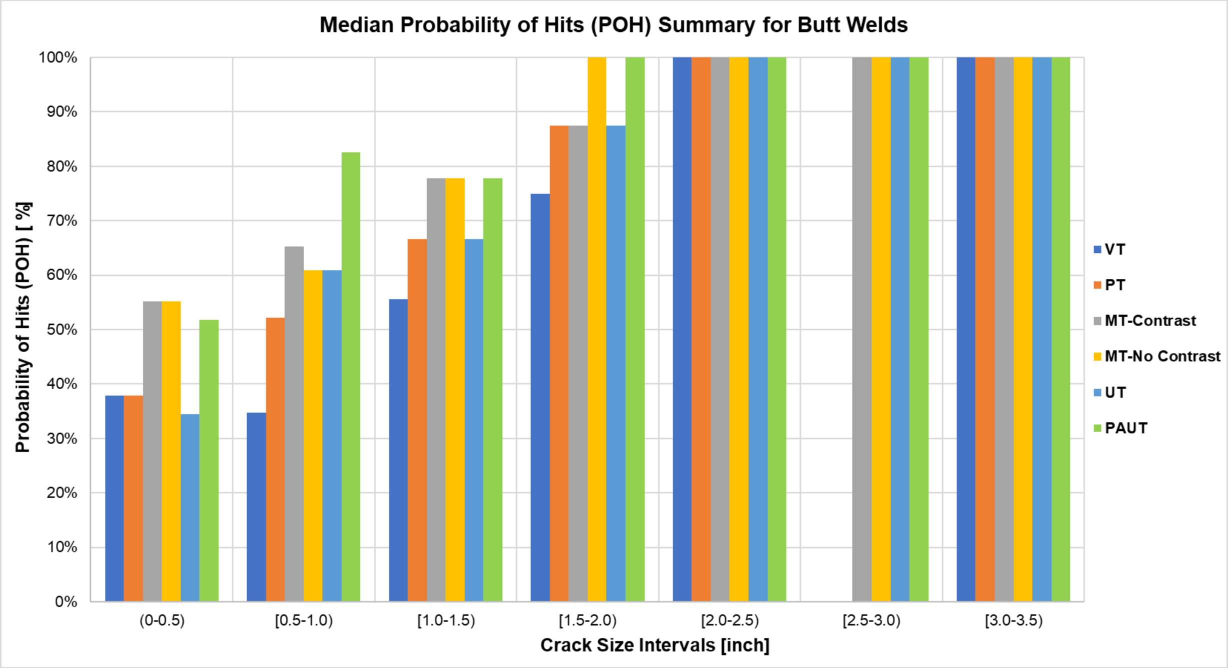


Figure 9. Median POH Summary for the BW

4.2 LOGIT/MLE

This section presents the POD results obtained while using the MLE based LOGIT approach. The MLE based method is a frequentist method that calculates a posteriori probability of the parameters of a model given the observations, by maximizing the likelihood of observing the data given the parameters. For this study, calculating the POD using the MLE method took place using NASA DOEPOD software.

4.2.1 FW

Figure 10 and Figure 11 show the summary median values of mean MLE POD and mean MLE POD 95 percent LCL for all operators that participated in each NDE method. As shown in Figure 10, MT without contrast and MT with contrast are the methods that achieved a POD greater than 0.9 for crack length over 0.93-inches. UT method achieved a POD greater than 0.9 for crack length over 1.6-inches approximately. VT and PT methods have the lowest POD for all crack lengths. Per the figures, the POD increases as the crack size increases for all inspection methods. Appendices E and F show the mean MLE POD and mean MLE POD at 95 percent confidence level for all inspectors and all NDE methods for FW panels.

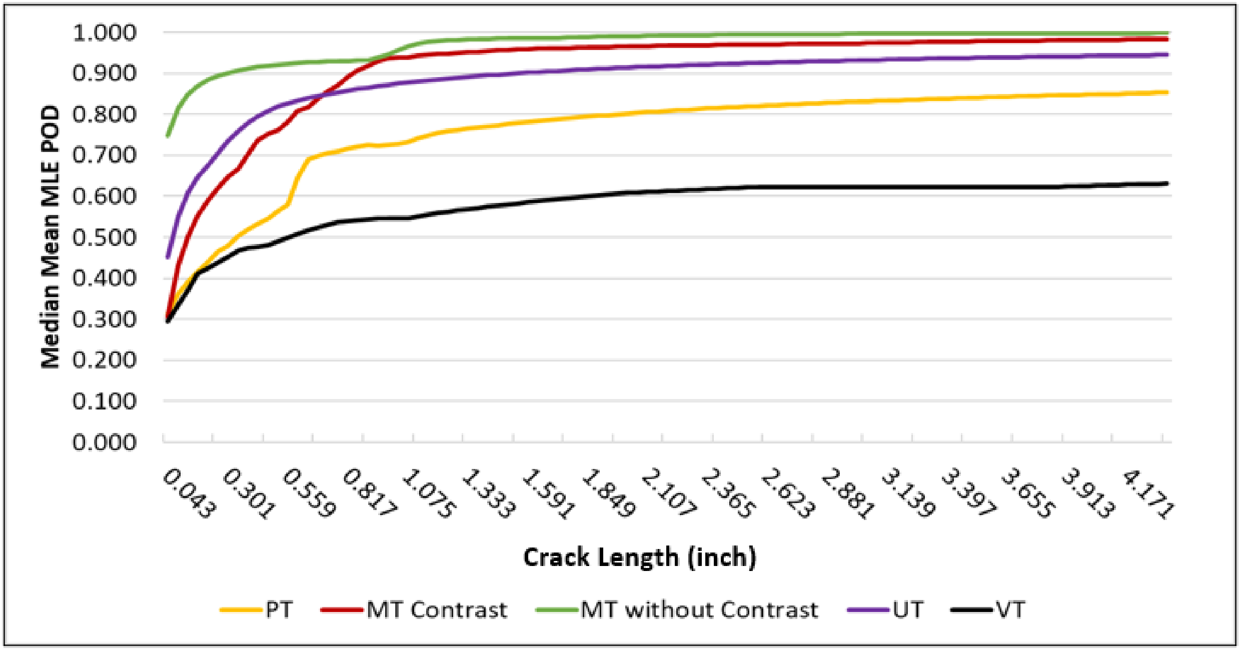


Figure 10. FW Median Summary Plot for Mean MLE POD

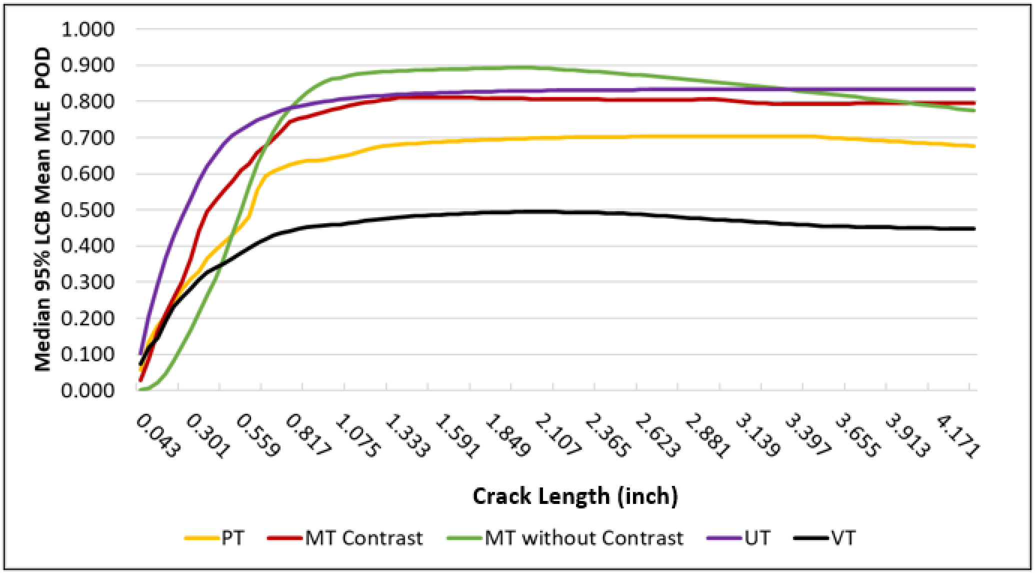


Figure 11. FW Median Summary Plot for Mean MLE POD with 95 percent LCL

4.2.2 BW

Figure 12 and Figure 13 show the summary median values of mean MLE POD and mean MLE POD 95 percent LCL for all operators that participated in each NDE method. In this case, PAUT method presents a POD higher than 90 percent for a larger range of crack length compared to the remaining methods. For both FW and BW panels it is important to review each operator's POD curves to understand the pattern of these curves, e.g., the number of operators that participated in each method is different as shown in Table 1. Appendices E and F show the mean POD and POD at 95 percent confidence level for all inspectors and NDE methods for BW panels.

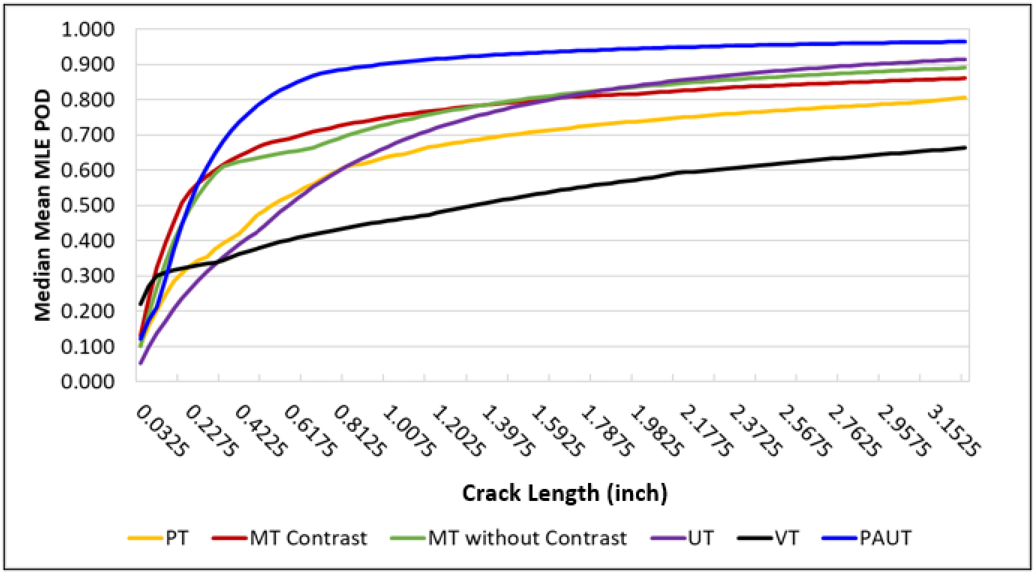


Figure 12. BW Median Summary Plot for Mean MLE POD

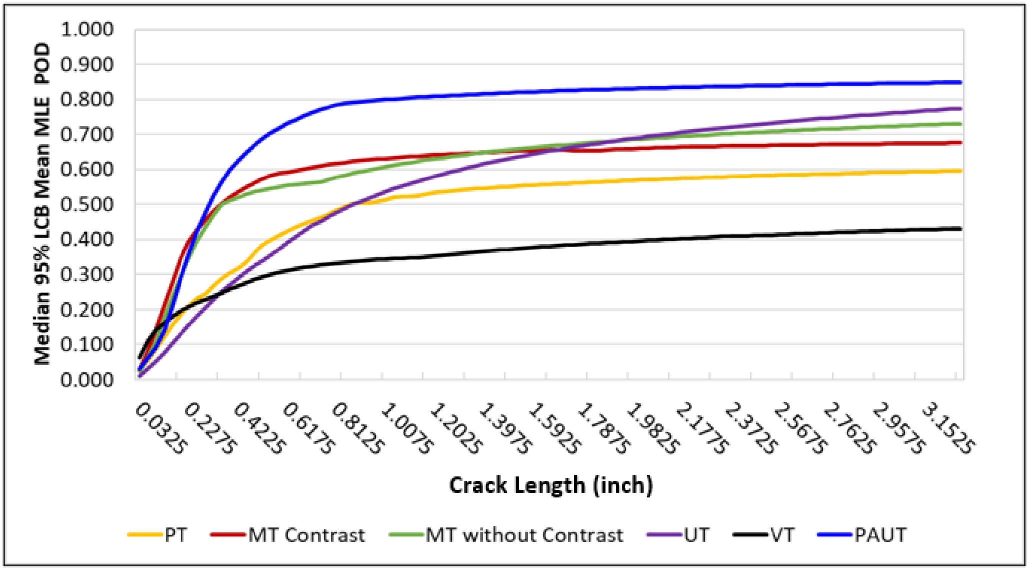
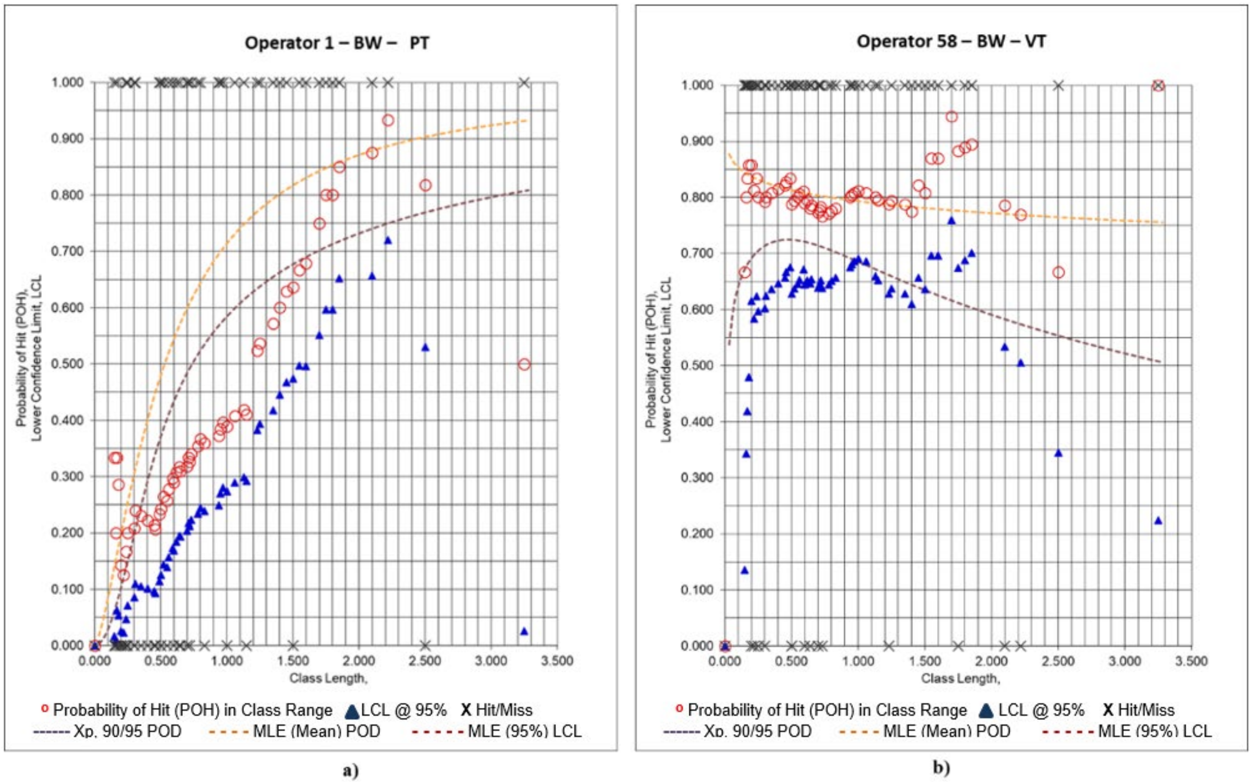


Figure 13. BW Median Summary Plot for Mean MLE POD with 95 Percent LCL

During TTCI's analysis, several cases were identified where the POD could not be calculated using the LOGIT/MLE approach.

Figure 14 shows a typical example of cases where this approach work and did not work. Figure 14(a) presents a typical example where researchers calculated the mean POD (i.e., yellow dashed lines) and the MLE at a 95 percent LCL (i.e., brown dashed lines) using the MLE approach. The MLE-LOGIT approach assumes that the typical curve that describes the mean POD monotonically increases as the crack length increases. However, there were several cases identified where the algorithm results did not show this pattern due to convergence issues (divergence) using this approach, which is shown in Figure 14(b).



a)

b)

Figure 14. Examples of MLE Convergence and Non-convergence

These findings agree with Generazio (2015) that states that the results from the MLE method may lack validity due to algorithm convergence and the inadequacy of the MLE method for NDE systems: “Use of MLE POD methods for fracture critical POD inspection demonstrations is not recommended due to the lack of validated NDE math models used in MLE.”

4.3 DOEPOD

Researchers based the DOEPOD on the 2-parameter binomial distribution where 2 possible outcomes are possible, hits or misses (e.g., 100 and 0). The DOEPOD method considers that cracks are not created equally, but they might be grouped by size, length, depth, etc., and it is a confidence interval-based approach. This DOEPOD approach also does not make assumptions of the POD model, as such it will not force the data to follow a specific curve (e.g., larger crack sizes will always have greater POD compared to smaller crack sizes).

False calls are also not directly associated with the DOEPOD POD results obtained like other approaches reported in this work, but DOEPOD allows to include it as a parameter in the template where the total number of false calls for each operator is recorded. This subsequently allow us to calculate the false calls percentage (FCP). Generally, an FCP exceeding 5 percent is unacceptable, as it indicates an excessively high scrap rate for good parts. AGARD (1993) suggested this 5 percent FCP for use with multi-parameter MLE curve fits. However, it is still not clear whether this suggestion was for LOGIT, PROBIT, or for the a/â curve fit approaches, and Wald or Likelihood ratio bounds. These are all important in deciding this acceptance level. DOEPOD yields a warning when the upper confidence bound (UCB) of the FCP exceeds 0.03448. The observed 90/95 POD results, when the UCB of the FCP exceeds 0.03448, it is not valid. The 3.448 percent is the quantitative upper Clopper-Pearson 95 percent bound at which the FCP may produce a “Lucky Hit” that is added to the number of hits, resulting in an erroneous higher estimate of the POD.

The research team analyzed railroad tank car NDE evaluation data results conducted from 1998 to 2016 using DOEPOD software. DOEPOD analyzed a total of 194 POD test datasets to yield a case identification for each dataset. Table 6 provides a comprehensive top-level summary of the DOEPOD analyses of the railroad tank car NDE inspection data. In addition, this table also provides DOEPOD recommendations to complete the validation over a range of larger flaw sizes.

Table 6. Comprehensive DOEPOD Summary of all Cases and Recommendation

CASE ID	Number of Datasets	DOEPOD Recommendations
CASE 1	16	Provide justification for false calls
CASE 1+	3	Provide justification for false calls
CASE 1#	6	Further validation at larger flaws. Add test specimens with larger flaws. Provide justification for false calls
CASE 1*	2	Further validation at larger flaws. Add test specimens with larger flaws. Provide justification for false calls
CASE 2	3	Add test specimens at identified flaw sizes to demonstrate POD to be monotonically increasing with flaw size
CASE 4	6	Increase amount of relevant data by adding test specimens at identified flaw sizes to establish acceptable POD
CASE 5	5	Add test specimens with increased flaw sizes to address excessive false negatives at smaller flaw sizes
CASE 6	78	Add test specimens with flaw sizes at least twice as large to address local inspection system oscillation instability or utilize a different inspection system or method
CASE 7	75	Add test specimens with flaw sizes at least twice as large to address global inspection system instability or utilize a different inspection system or method

CASE 1, CASE 1#, CASE 1*, and CASE 2 all exhibit at least one point where the one-sided lower 95 percent confidence bound on POD exceeds 0.90 at a given flaw size. CASE 4 represents the datasets that are identical to CASE 2; nonetheless, researchers needed additional data results on selected flaw sizes to move a CASE 4 to a CASE 2 dataset. Similarly, CASE 6 datasets exhibit local instability over a portion of the flaw sizes tested; therefore, there is a need for data results for larger flaw sizes or the inspection system is inappropriate for the inspection required. Finally, CASE 7 datasets exhibit instability over the entire the flaw size range tested; therefore, data results for larger flaw sizes are needed or the inspection system is inappropriate for the inspection required.

Table 6, Table 7, and Table 8 presented the breakdown of the data, while Table 6 further elaborated the BW and FW, respectively. From these tables, several of the DOEPOD analysis for the BW shows that the NDE system suffered from local/global instability. DOEPOD also recommends adding test specimens with flaw sizes at least twice as large as the best LCL to address local/global inspection system oscillation instability or utilize a different NDE inspection system or method to achieve 90/95 POD.

Similarly, DOEPOD analysis for the FW shows mixed results with CASE 6 and CASE 7. Some of the operators could demonstrate the good case (i.e., CASES 1, 1#, 1*, and 2) for MT inspection for FW, but most of them suffered from rejectable FCP. DOEPOD once again recommends adding test specimens with flaw sizes at least twice as large as the best LCL to address local/global inspection system oscillation instability or utilize a different NDE inspection system or method to achieve 90/95 POD.

Table 7. Summary of DOEPOD Cases for FW

CASES	FW					
	VT	PT	MT with CONTRAST	MT without CONTRAST	UT	TOTAL
CASE 1	0	3	7	6	0	16
CASE 1+	1	0	2	0	0	3
CASE 1 #	0	1	3	1	0	5
CASE 1*	0	0	0	2	0	2
CASE 2	0	1	1	1	0	3
CASE 4	1	0	2	0	0	3
CASE 5	0	0	0	0	0	0
CASE 6	10	6	6	0	3	25
CASE 7	14	16	3	0	0	33
Total Count	**26**	**27**	**24**	**10**	**3**	**90**

Table 8. Summary of DOEPOD Cases for BW

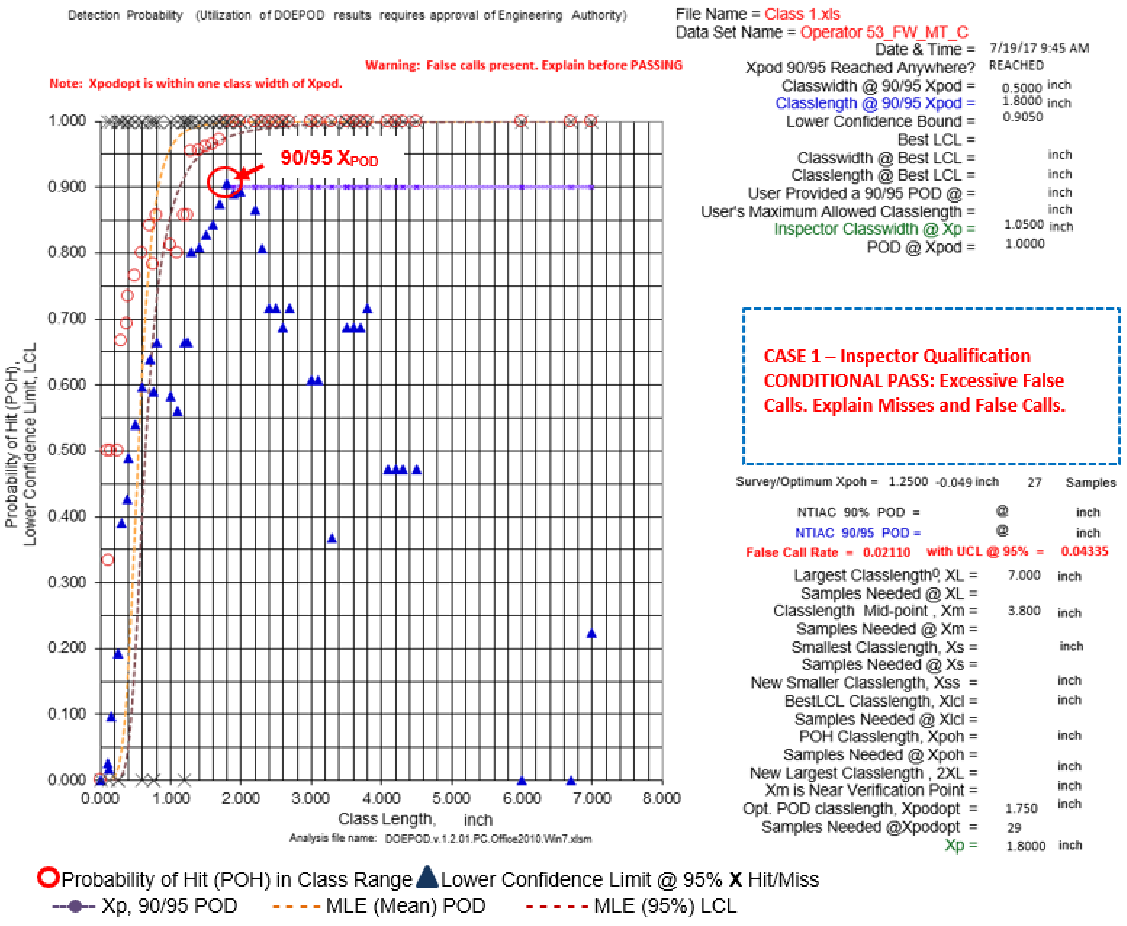
CASES	BW						
	VT	PT	MT with CONTRAST	MT without CONTRAST	UT	PAUT	TOTAL
CASE 1	0	0	0	0	0	0	0
CASE 1+	0	0	0	0	0	0	0
CASE 1 #	0	0	0	1	0	0	1
CASE 1*	0	0	0	0	0	0	0
CASE 2	0	0	0	0	0	0	0
CASE 4	0	0	1	1	1	0	3
CASE 5	0	1	0	3	0	1	5
CASE 6	16	10	9	2	15	1	53
CASE 7	7	14	7	4	9	1	42
Total Count	**23**	**25**	**17**	**11**	**25**	**3**	**104**

In addition to obtaining 90/95 POD results for validating NDE inspection systems, DOEPOD analysis can be used for evaluating the qualification of NDE inspectors. The 90/95 POD capability must be demonstrated first, by obtaining CASE 1 or CASE 1+ with NDE inspection processes and procedures fixed and under control, before asking inspectors to demonstrate their inspection capability using the inspection system (U.S. Department of Defense, 2004). Since in most cases 90/95 POD was not demonstrated and, in some cases, where 90/95 was demonstrated, it suffered from high FCP; therefore, NDE inspector qualification was not performed and is not presented in this report.

Next, the DOEPOD cases described above are briefly illustrated for the reader's convenience. For this, the following sections presented the MT with contrast results. Appendix G and Appendix H details all the POD test datasets analyzed using DOEPOD.

CASE 1

Figure 15 shows that this is the best-case scenario. It suggests that there is an adequate distribution of flaws at X_{POD}, and there are enough well-distributed large flaws above the X_{POD} flaw size. A 90/95 POD is reached at a flaw size of 1.8-inch and there are no misses above X_{POD}. However, the 95 percent one-sided UCB FCP or the "probability of false call" is 0.04335 (4.3%). When the UCB of the FCP exceeds 3.4 percent, the POD is not valid. Therefore, DOEPOD recommends validating 90/95 POD from X_{POD} to the largest flaw, X_L.



File Name = Class 1.xls
Data Set Name = Operator 53_FW_MT_C
Date & Time = 7/19/17 9:45 AM
Xpod 90/95 Reached Anywhere? REACHED
Classwidth @ 90/95 Xpod = 0.5000 inch
Classlength @ 90/95 Xpod = 1.8000 inch
Lower Confidence Bound = 0.9050
Best LCL =
Classwidth @ Best LCL = inch
Classlength @ Best LCL = inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = 1.0500 inch
POD @ Xpod = 1.0000

CASE 1 – Inspector Qualification
CONDITIONAL PASS: Excessive False Calls. Explain Misses and False Calls.

Survey/Optimum Xpoh = 1.2500 -0.049 inch 27 Samples

NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.02110 with UCL @ 95% = 0.04335

Largest Classlength⁰, XL = 7.000 inch
Samples Needed @ XL =
Classlength Mid-point , Xm = 3.800 inch
Samples Needed @ Xm =
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = inch
Samples Needed @ Xlcl =
POH Classlength, Xpoh = inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = 1.750 inch
Samples Needed @Xpodopt = 29
Xp = 1.8000 inch

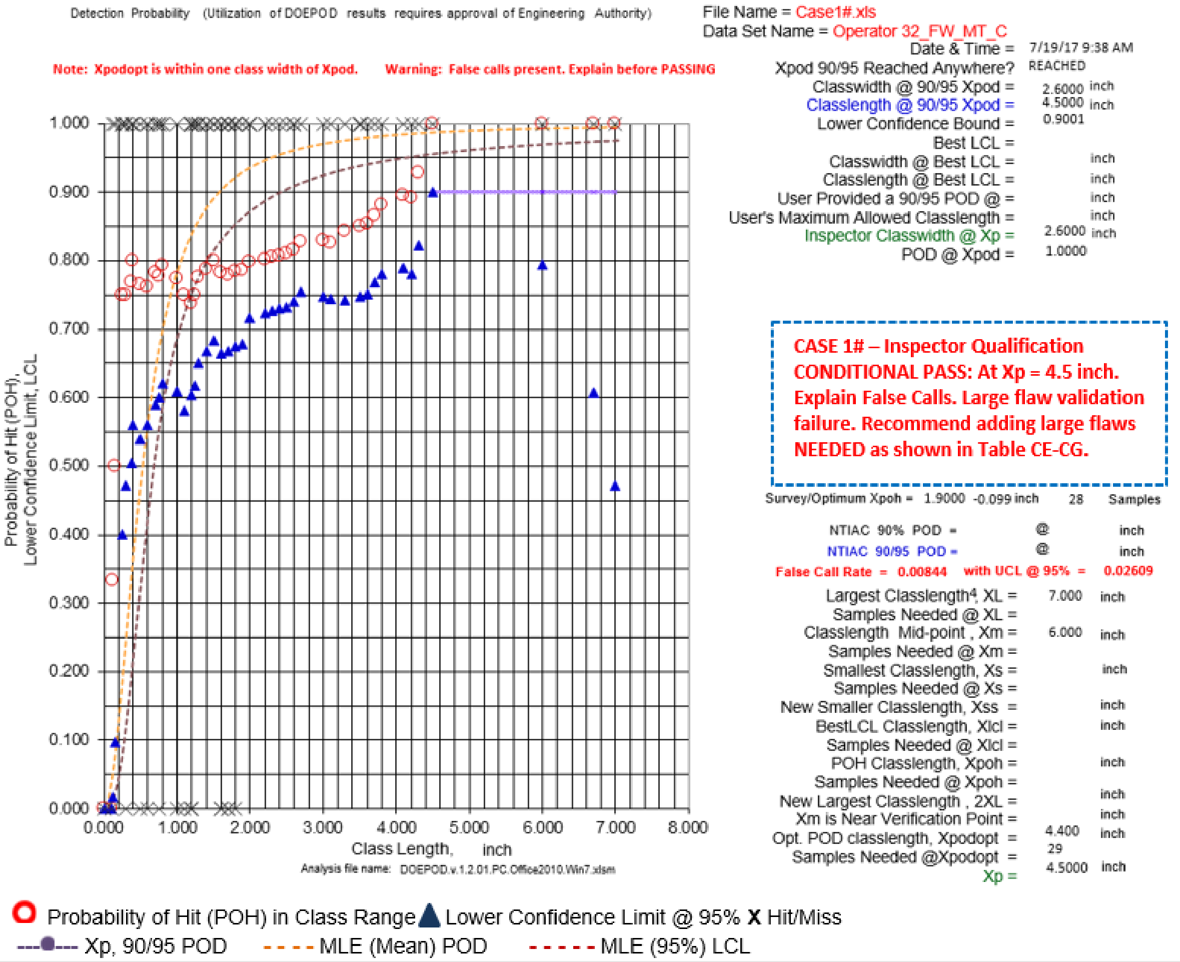
Figure 15. CASE 1 Example of DOEPOD Analysis for Operator 53 (MT with contrast)

CASE 1+

None obtained.

CASE 1#

Figure 16 shows CASE 1#. A 90/95 POD is reached at a flaw size 4.5-inch, and there are misses only below a flaw size of 2.0-inch. There was an acceptable number of false calls with a 95 percent one-sided UCB of 0.02609 (2.6%). The POD requires further validation to verify if it is increasing with increasing class length. The DOEPOD recommendations are to add the specified large flaws (4.4-inch) and explain the false calls.



File Name = Case1#.xls
Data Set Name = Operator 32_FW_MT_C
Date & Time = 7/19/17 9:38 AM
Xpod 90/95 Reached Anywhere? REACHED
Classwidth @ 90/95 Xpod = 2.6000 inch
Classlength @ 90/95 Xpod = 4.5000 inch
Lower Confidence Bound = 0.9001
Best LCL =
Classwidth @ Best LCL = inch
Classlength @ Best LCL = inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = 2.6000 inch
POD @ Xpod = 1.0000

CASE 1# – Inspector Qualification
CONDITIONAL PASS: At Xp = 4.5 inch.
Explain False Calls. Large flaw validation failure. Recommend adding large flaws NEEDED as shown in Table CE-CG.

Survey/Optimum Xpoh = 1.9000 -0.099 inch 28 Samples

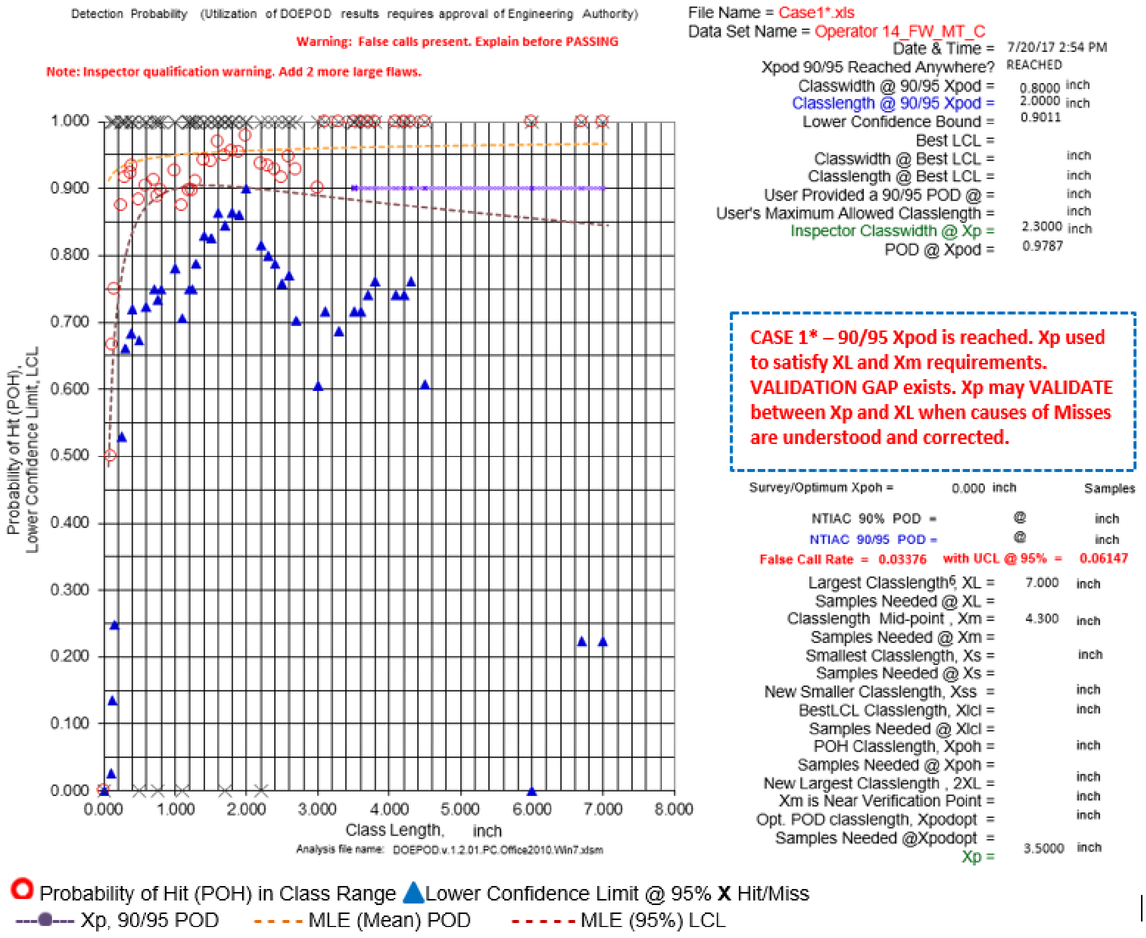
NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.00844 with UCL @ 95% = 0.02609

Largest Classlength^4, XL = 7.000 inch
Samples Needed @ XL =
Classlength Mid-point , Xm = 6.000 inch
Samples Needed @ Xm =
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = inch
Samples Needed @ Xlcl =
POH Classlength, Xpoh = inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = 4.400 inch
Samples Needed @Xpodopt = 29
Xp = 4.5000 inch

Figure 16. CASE 1# Example of DOEPOD Analysis for Operator 32 (MT with contrast)

CASE 1*

Figure 17 shows CASE 1*. A 90/95 POD is reached at a flaw size of 2.0 inches, and there is one miss above XPOD at a flaw size of 2.2-inch. Also, the POH is fluctuating within the class range 2.2-inch to 3.0-inch. There was an unacceptable number of false calls with a 95 percent one-sided UCB of 0.06147 (6.1%). Since the FCP exceeds the accepted threshold of 3.4 percent, the POD is not valid. Therefore, DOEPOD recommends validating 90/95 POD from XPOD to the largest flaw, X_L because a validation gap may exist between X_L and X_m. However, X_p may be used to validate the 90/95 POD between X_p and X_L only when causes of misses are understood and corrected above X_{POD}.



Detection Probability (Utilization of DOEPOD results requires approval of Engineering Authority)

Warning: False calls present. Explain before PASSING

Note: Inspector qualification warning. Add 2 more large flaws.

Analysis file name: DOEPOD.v.1.2.01.PC.Office2010.Win7.xlsm

File Name = Case1*.xls
Data Set Name = Operator 14_FW_MT_C
Date & Time = 7/20/17 2:54 PM
Xpod 90/95 Reached Anywhere? REACHED
Classwidth @ 90/95 Xpod = 0.8000 inch
Classlength @ 90/95 Xpod = 2.0000 inch
Lower Confidence Bound = 0.9011
Best LCL =
Classwidth @ Best LCL = inch
Classlength @ Best LCL = inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = 2.3000 inch
POD @ Xpod = 0.9787

CASE 1* – 90/95 Xpod is reached. Xp used to satisfy XL and Xm requirements. VALIDATION GAP exists. Xp may VALIDATE between Xp and XL when causes of Misses are understood and corrected.

Survey/Optimum Xpoh = 0.000 inch Samples

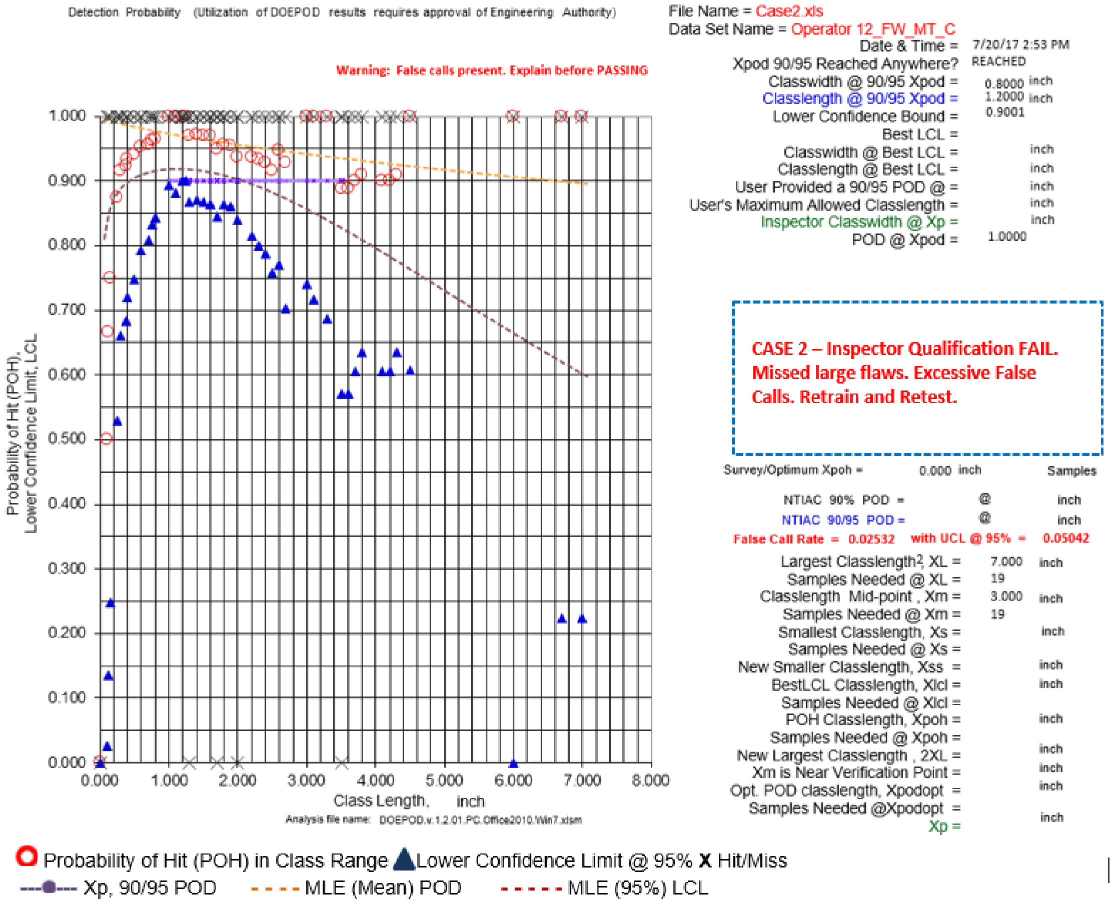
NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.03376 with UCL @ 95% = 0.06147

Largest Classlength⁶, XL = 7.000 inch
Samples Needed @ XL =
Classlength Mid-point , Xm = 4.300 inch
Samples Needed @ Xm =
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = inch
Samples Needed @ Xlcl =
POH Classlength, Xpoh = inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = inch
Samples Needed @Xpodopt =
Xp = 3.5000 inch

Figure 17. CASE 1* Example of DOEPOD Analysis for Operator 14 (MT with contrast)

CASE 2

Figure 18 shows CASE 2. A 90/95 POD is reached at a flaw size 1.2-inch, and there are misses above X_{POD}. An explanation and resolution of all misses above X_{POD} are required. Also, the POH is fluctuating within the class range 1.2-inch to 4.3-inch. There was an unacceptable number of false calls with a 95 percent one-sided UCB of 0.05042 (5.0%). Since the FCP exceeds the accepted threshold of 3.4 percent, the POD is not valid. Therefore, DOEPOD recommends validating 90/95 POD at identified flaw sizes. Also, the false calls need to be addressed and corrected.



File Name = Case2.xls
Data Set Name = Operator 12_FW_MT_C
Date & Time = 7/20/17 2:53 PM
Xpod 90/95 Reached Anywhere? REACHED
Classwidth @ 90/95 Xpod = 0.8000 inch
Classlength @ 90/95 Xpod = 1.2000 inch
Lower Confidence Bound = 0.9001
Best LCL =
Classwidth @ Best LCL = inch
Classlength @ Best LCL = inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = inch
POD @ Xpod = 1.0000

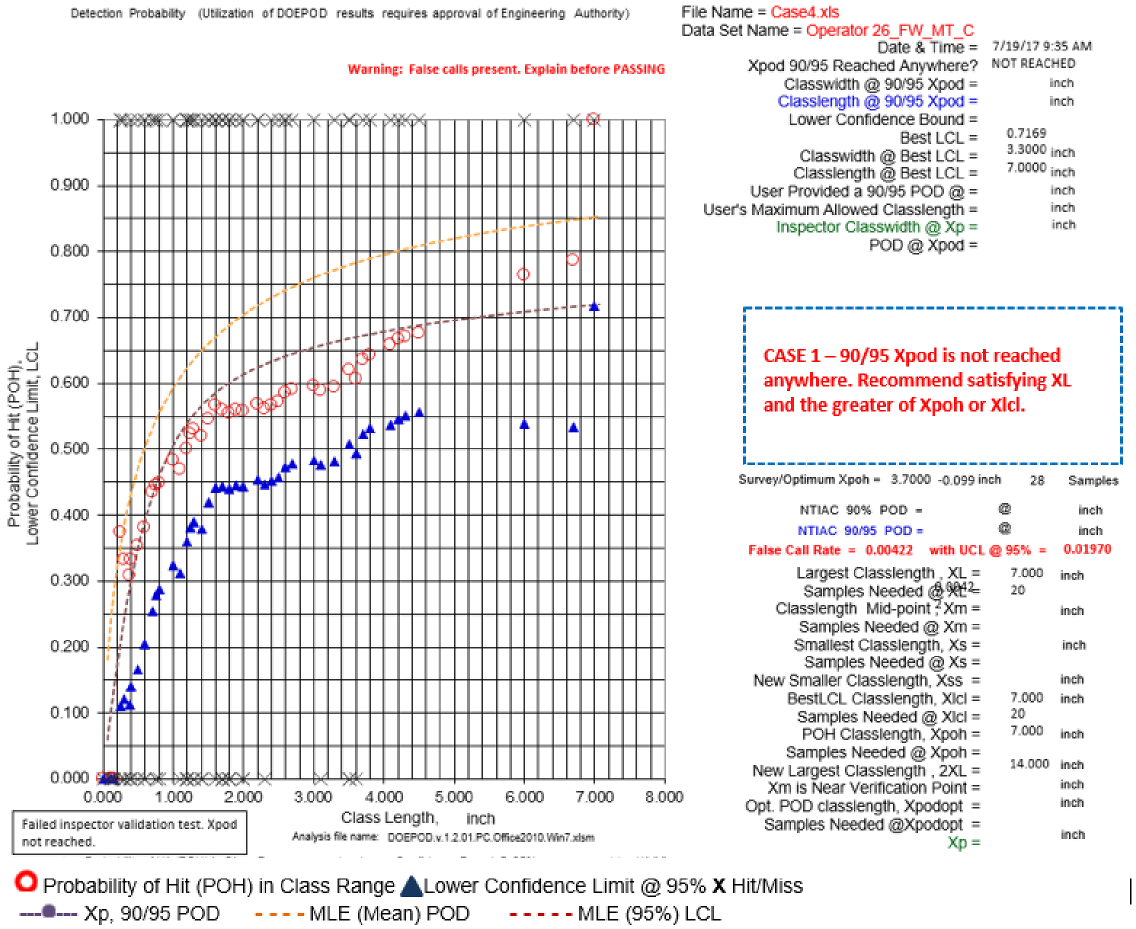
CASE 2 – Inspector Qualification FAIL. Missed large flaws. Excessive False Calls. Retrain and Retest.

Survey/Optimum Xpoh = 0.000 inch Samples
NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.02532 with UCL @ 95% = 0.05042
Largest Classlength2, XL = 7.000 inch
Samples Needed @ XL = 19
Classlength Mid-point , Xm = 3.000 inch
Samples Needed @ Xm = 19
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = inch
Samples Needed @ Xlcl =
POH Classlength, Xpoh = inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = inch
Samples Needed @Xpodopt =
Xp = inch

Figure 18. CASE 2 Example of DOEPOD Analysis for Operator 12 (MT with contrast)

CASE 4

CASE 4 is like CASE 2, except that 90/95 POD at X_{POD} is not reached anywhere, as shown in Figure 19. This graph shows an estimated POD at 7.0-inch flaw size is 1 (100 percent) with a 95 percent one-sided LCL of 0.7169 (71.7%). There are no misses at or greater than the $X_{Best\ LCL}$ class length, or within the class width group exhibiting the best LCL, $X_{Best\ LCL}$. DOEPOD recommends satisfying X_L and the greater of X_{POH} or X_{LCL}.


Detection Probability (Utilization of DOEPOD results requires approval of Engineering Authority)
Warning: False calls present. Explain before PASSING
Probability of Hit (POH), Lower Confidence Limit, LCL
1.000
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
0.000
1.000
2.000
3.000
4.000
5.000
6.000
7.000
8.000
Class Length, inch
Probability of Hit (POH) in Class Range
Lower Confidence Limit @ 95%
X Hit/Miss
Xp, 90/95 POD
MLE (Mean) POD
MLE (95%) LCL

Failed inspector validation test. Xpod not reached.

Analysis file name: DOEPOD.v.1.2.01.PC.Office2010.Win7.xlsm

File Name = Case4.xls
Data Set Name = Operator 26_FW_MT_C
Date & Time = 7/19/17 9:35 AM
Xpod 90/95 Reached Anywhere? NOT REACHED
Classwidth @ 90/95 Xpod = inch
Classlength @ 90/95 Xpod = inch
Lower Confidence Bound =
Best LCL = 0.7169
Classwidth @ Best LCL = 3.3000 inch
Classlength @ Best LCL = 7.0000 inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = inch
POD @ Xpod =

CASE 1 – 90/95 Xpod is not reached anywhere. Recommend satisfying XL and the greater of Xpoh or Xlcl.

Survey/Optimum Xpoh = 3.7000 -0.099 inch 28 Samples

NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.00422 with UCL @ 95% = 0.01970

Largest Classlength , XL = 7.000 inch
Samples Needed @ XL = 20
Classlength Mid-point , Xm = inch
Samples Needed @ Xm =
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = 7.000 inch
Samples Needed @ Xlcl = 20
POH Classlength, Xpoh = 7.000 inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = 14.000 inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = inch
Samples Needed @Xpodopt =
Xp = inch

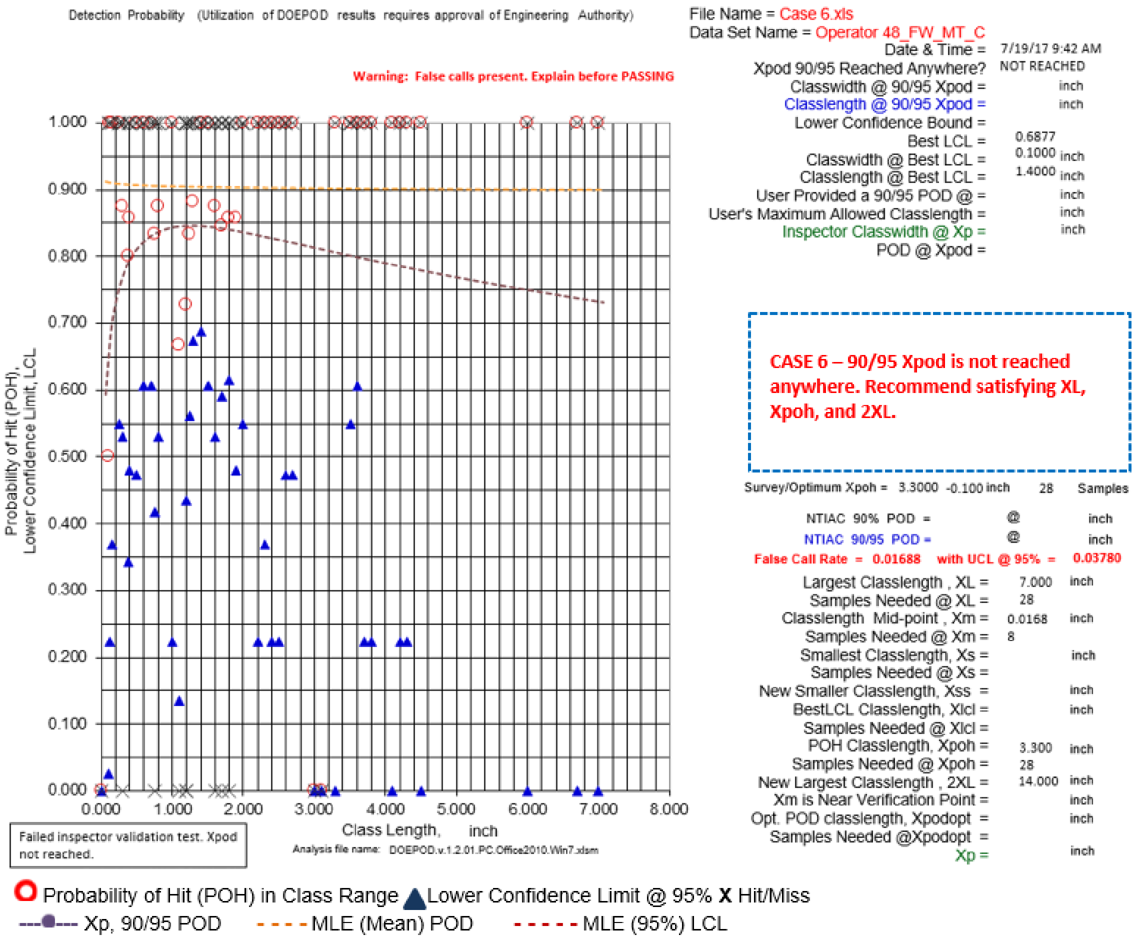
Figure 19. CASE 4 Example of DOEPOD Analysis for Operator 26 (MT with contrast)

CASE 5

None obtained.

CASE 6

A 90/95 POD at X_{POD} has not reached anywhere, as shown in Figure 20. This graph shows an estimated POD at 1.4-inch flaw size is 0.90 (90 percent) with a 95 percent one-sided LCL of 0.6877 (68.7%). There are misses above $X_{Best\ LCL}$, which require an explanation. There exists a class length (3.3-inches), X_{POH}=1, above which there are no misses. The DOEPOD recommendations are to satisfy 90/95 POD at X_L, X_{POH}, and $2X_L$, respectively.



```
File Name = Case 6.xls
Data Set Name = Operator 48_FW_MT_C
                    Date & Time =  7/19/17 9:42 AM
   Xpod 90/95 Reached Anywhere?  NOT REACHED
       Classwidth @ 90/95 Xpod =              inch
      Classlength @ 90/95 Xpod =              inch
      Lower Confidence Bound =
                    Best LCL =     0.6877
       Classwidth @ Best LCL =     0.1000 inch
      Classlength @ Best LCL =     1.4000 inch
    User Provided a 90/95 POD @ =             inch
User's Maximum Allowed Classlength =          inch
      Inspector Classwidth @ Xp =             inch
                 POD @ Xpod =
```

CASE 6 – 90/95 Xpod is not reached anywhere. Recommend satisfying XL, Xpoh, and 2XL.

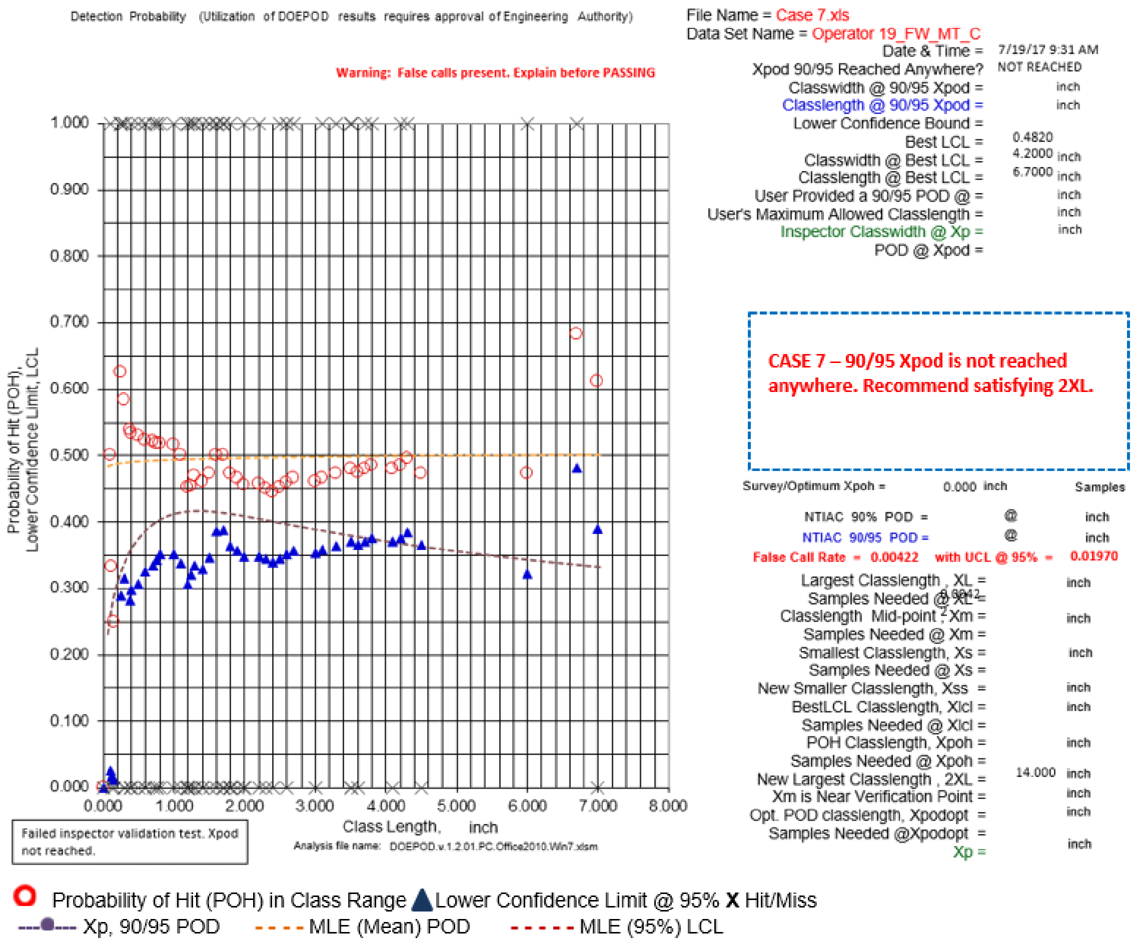
```
Survey/Optimum Xpoh =  3.3000  -0.100 inch    28    Samples

      NTIAC  90%  POD  =             @          inch
      NTIAC  90/95  POD =            @          inch
False Call Rate  =  0.01688   with UCL @ 95%  =  0.03780
       Largest Classlength , XL =    7.000   inch
          Samples Needed @ XL =     28
     Classlength  Mid-point , Xm =  0.0168   inch
          Samples Needed @ Xm =  8
      Smallest Classlength, Xs =             inch
          Samples Needed @ Xs =
   New Smaller Classlength, Xss  =           inch
      BestLCL Classlength, Xlcl =            inch
        Samples Needed @ Xlcl =
        POH Classlength, Xpoh =     3.300    inch
        Samples Needed @ Xpoh =     28
   New Largest Classlength , 2XL =  14.000   inch
   Xm is Near Verification Point =           inch
  Opt. POD classlength, Xpodopt  =           inch
    Samples Needed @Xpodopt  =
                          Xp =               inch
```

Figure 20. CASE 6 Example of DOEPOD Analysis for Operator 48 (MT with contrast)

CASE 7

CASE 7 is like CASE 6, a 90/95 POD at XPOD is not reached anywhere, as shown in Figure 21. This graph shows an estimated POD at 0.5-inch flaw size is 0.491 (49.1%). However, the best 95 percent one-sided LCL POD is demonstrated at 6.7-inches and equals to 0.482 (48.2 percent). The POH (red circles) is fluctuating throughout the entire range of flaw sizes used; and therefore, the inspection procedure or inspection system is inadequate for flaws 7-inches or less. In addition, there are excessive misses above and below $X_{\text{Best LCL}}$. There does not exist a class length, $X_{\text{POH=1}}$, above which there are no misses. Similarly, MLE-LOGIT curve fit method for estimating POD also failed with a divergence warning and the curve fit shown cannot be used in this case. Since the POD has not been validated to increase with flaw size, then the inspection procedure or inspection system is inadequate for flaws 7-inch or less, and either additional training is needed or the inspection system may not be appropriate for meeting inspection criteria, or there is a need to expand the current range of X_L by adding 29 new samples with class lengths of $2X_L$ or greater.


Detection Probability (Utilization of DOEPOD results requires approval of Engineering Authority)
Warning: False calls present. Explain before PASSING
Probability of Hit (POH), Lower Confidence Limit, LCL
1.000
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
0.000
1.000
2.000
3.000
4.000
5.000
6.000
7.000
8.000
Class Length, inch
Analysis file name: DOEPOD.v.1.2.01.PC.Office2010.Win7.xlsm
Failed inspector validation test. Xpod not reached.
Probability of Hit (POH) in Class Range
Lower Confidence Limit @ 95%
X Hit/Miss
Xp, 90/95 POD
MLE (Mean) POD
MLE (95%) LCL

File Name = Case 7.xls
Data Set Name = Operator 19_FW_MT_C
Date & Time = 7/19/17 9:31 AM
Xpod 90/95 Reached Anywhere? NOT REACHED
Classwidth @ 90/95 Xpod = inch
Classlength @ 90/95 Xpod = inch
Lower Confidence Bound =
Best LCL = 0.4820
Classwidth @ Best LCL = 4.2000 inch
Classlength @ Best LCL = 6.7000 inch
User Provided a 90/95 POD @ = inch
User's Maximum Allowed Classlength = inch
Inspector Classwidth @ Xp = inch
POD @ Xpod =

CASE 7 – 90/95 Xpod is not reached anywhere. Recommend satisfying 2XL.

Survey/Optimum Xpoh = 0.000 inch Samples

NTIAC 90% POD = @ inch
NTIAC 90/95 POD = @ inch
False Call Rate = 0.00422 with UCL @ 95% = 0.01970

Largest Classlength , XL = inch
Samples Needed @ XL = 0.0042
Classlength Mid-point , Xm = inch
Samples Needed @ Xm =
Smallest Classlength, Xs = inch
Samples Needed @ Xs =
New Smaller Classlength, Xss = inch
BestLCL Classlength, Xlcl = inch
Samples Needed @ Xlcl =
POH Classlength, Xpoh = inch
Samples Needed @ Xpoh =
New Largest Classlength , 2XL = 14.000 inch
Xm is Near Verification Point = inch
Opt. POD classlength, Xpodopt = inch
Samples Needed @Xpodopt =
Xp = inch

Figure 21. CASE 7 Example of DOEPOD Analysis for Operator 19 (MT with contrast)

Finally, the research team completed a high-level comparison to compare DOEPOD results with the results obtained using MLE two-parameter logit model (LOGIT/MLE). Table 9 and Table 10 show the summary of the frequency of operators that reached the 90/95 POD and 90 percent POD for the DOEPOD and MLE methods for FW and BW panels, respectively. The tables also show that operators exceeded the number of the probability of FCP threshold of 3.448 percent from the DOEPOD software. A large percentage of operators exceeded the probability of false calls for both FW and BW. It is also demonstrated that the NDE methods and procedure used on BW failed to reach 90/95 POD with both software packages. For FW, differences between the two software results are seen when comparing the percentage of NDE operators that reached 90 XPOD for the two different methods. This once again demonstrates that the MLE-LOGIT 90/95 POD criteria may or may not be adequate for NDE system validation on tank car fusion welded components. Also, from the results it can be observed that the DOEPOD method tends to be more conservative than the MLE method because of the number of operators that reached the 90/95 POD.

In addition, it can be observed that BW panels have a lower percentage of operators that reached 90/95 XPOD and 90 XPOD MLE compared to the FW panels. Researchers considered two criteria for the BW panels' evaluation: (1) the length, and (2) the location of the cracks with an associated tolerance of +- 0.5-inches. The location criteria, apart from considering the distance to the initiation of the crack from the reference point, also included the side of the panel detected on the crack.

Table 9. Summary Results for DOEPOD and MLE Methods for FW

NDE Method	TOTAL Operators	Operators that Reached 90/95 XPOD (DOEPOD)	Operators that Reached 90 XPOD (MLE)	Datasets with Probability of False Calls that Exceed 3.448%
VT	26	3.85%	11.54%	92.31%
MT with contrast	24	54.17%	87.50%	91.67%
MT without contrast	10	100%	100%	100%
PT	27	18.52%	33.33%	55.56%
UT	3	0%	100%	100%

Table 10. Summary Results for DOEPOD and MLE Methods for BW

NDE Method	Total Operators	Operators that Reached 90/95 XPOD (DOEPOD)	Operators that Reached 90 XPOD (MLE)	Datasets with Probability of False Calls that Exceed 3.448%
VT	23	0%	4.35%	95.65%
MT with contrast	17	0%	41.18%	94.12%
MT without contrast	11	9.09%	45.45%	100%
PT	25	0%	16%	100%
UT	25	0%	48%	100%
PAUT	3	0%	66.67%	100%

Appendices G and H present the DOEPOD summary tables for each NDE method, operator on FW and BW panels, as well as the count of hits, misses, and false calls. The results for each operator also contain information on whether the operator reached 90/95 XPOD and the recommendations in terms of increasing the sample size to recalculate the POD to obtain the 90/95 XPOD. Additionally, appendices C and D show the DOEPOD output plots at the operator level.

For reference, Figure 22 presents an output for operator 58 using the PT method on FW panels. The x-axis shows the crack length class which is an interval defined by the DOEPOD software that accounts for small increments of the crack length for the calculation. The y-axis represents the POD or POH, per the DOEPOD software. The various curves in the plot contain information about the POD in the crack length class range using the DOEPOD method, the LCL at 95 percent using the DOEPOD method, the mean POD using the MLE method, the LCL at 95 percent using the MLE method, hits and misses, and whether 90/95 POD or greater is achieved. As shown in Figure 22, the operator did not reach the 90/95 POD, and the POD and LCL at 95 percent curves using the DOEPOD method does not show a monotonically increasing pattern. Instead, it can be observed that the POD for crack length class ranges of for example, 3-inch, is smaller than POD of a crack length of any crack length of less than 3-inches.

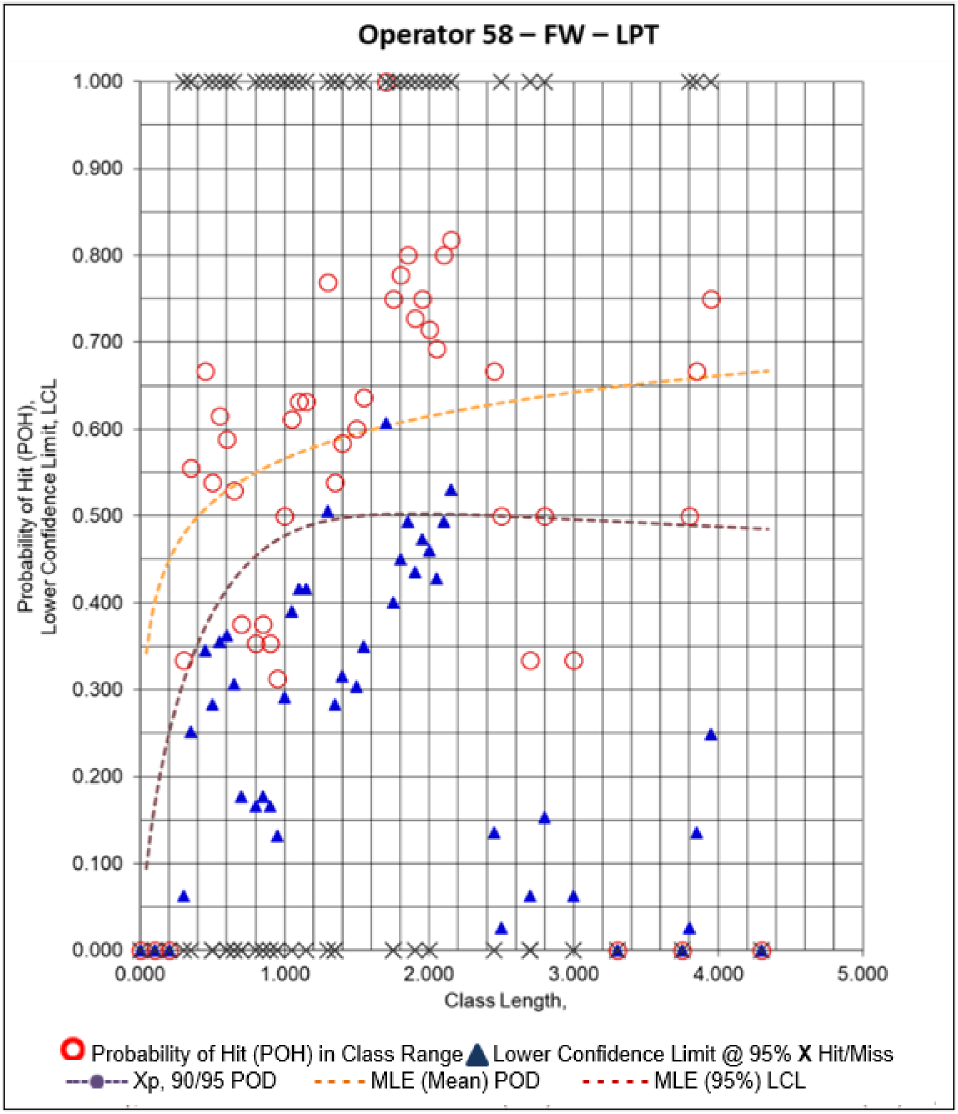


Figure 22. DOEPOD Results for FW Using PT – Operator 58

Figure 23 to Figure 26 show the summary POD and LCL POD at 95 percent using the DOEPOD method for FW and BW for each inspection method. The summary plots are based on the median value. Although, researchers recommended looking at the results for each operator and inspection method, these figures provide an initial description of the POD for each inspection method.

The POD and LCL POD at 95 percent did not show a monotonically increasing pattern for some of the inspection methods. This means that the POD varied depending on the crack length but not in an increasing fashion, i.e., a high POD might not be related to a large crack length.

For FW, the MT without contrast method POD showed an overall increasing pattern based on the median POD values and exceeded the 90 percent from a crack length class of 0.3-inches forward. The median LCL at 95 percent did not show an increasing pattern where in some cases the LCL at 95 percent was lower for large crack length for class compared to smaller crack length classes. This means that for example, the LCL for POD at 95 percent for a 3.3-inch crack length class was lower than the LCL for a 0.3-inch crack length class. MT with contrast showed a similar behavior to MT without contrast. The median POD values for PT and UT methods did not show a constant increasing pattern and did not reach 90 percent POD.

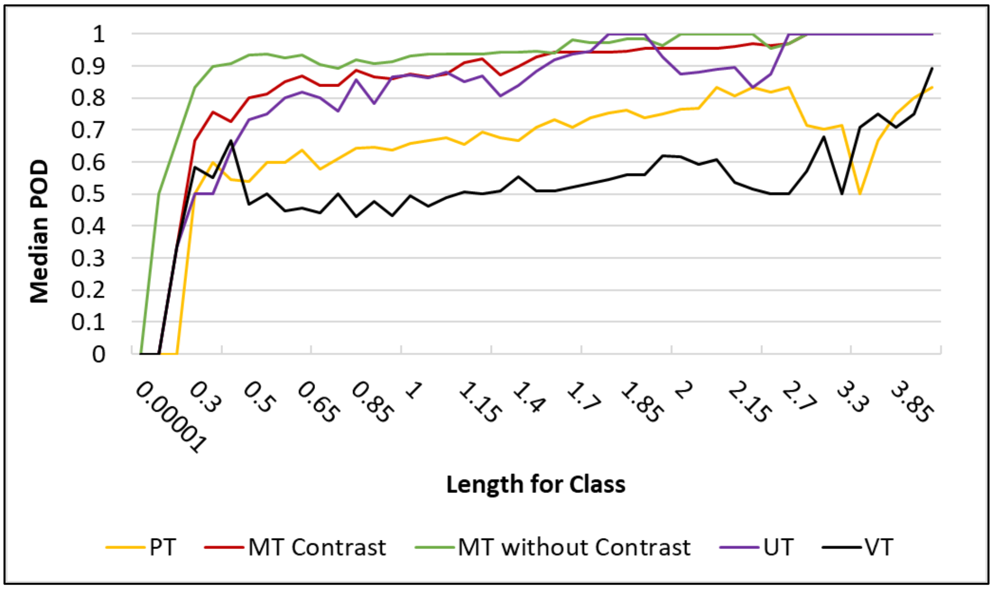


Figure 23. Summary Plot Median POD – FW

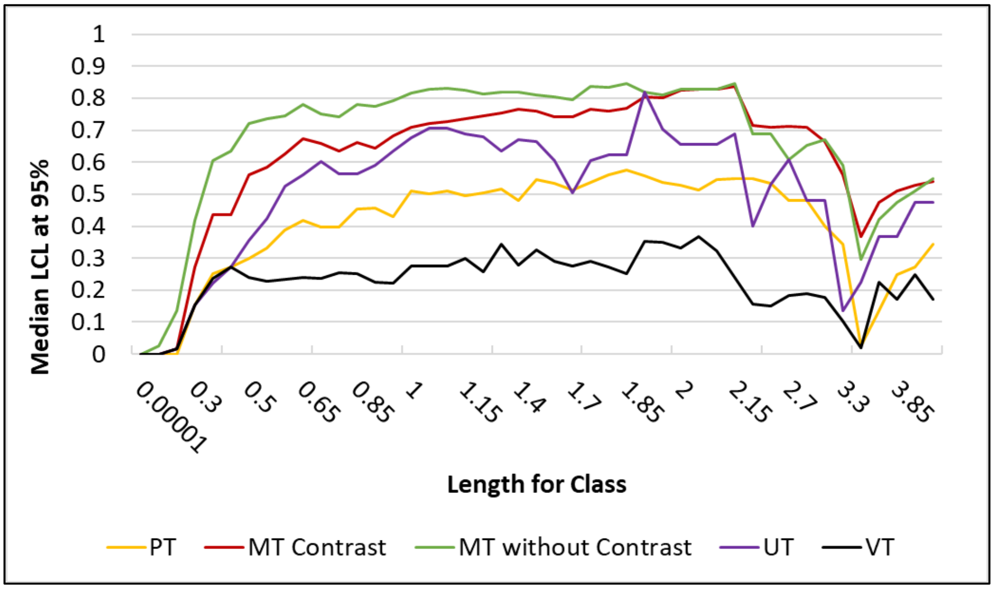


Figure 24. Summary Plot Median LCL at 95 Percent POD – FW

For BW, the median POD for all methods seemed to reach 90 percent for all methods as shown in Figure 25 and Figure 26. PAUT is the method that showed the largest changes of median POD compared to the remaining methods. It is important to point out that these plots in Figure 25 and Figure 26 provide a general idea of the POD and its LCL for each method, and that there is a need for an analysis per operator to understand the operator's POD per crack length class.

In terms of the median LCL at 95 percent, none of the methods reached 90 percent POD. For all methods, there were cases where higher crack length classes reported a lower LCL at 95 percent than smaller crack length classes.

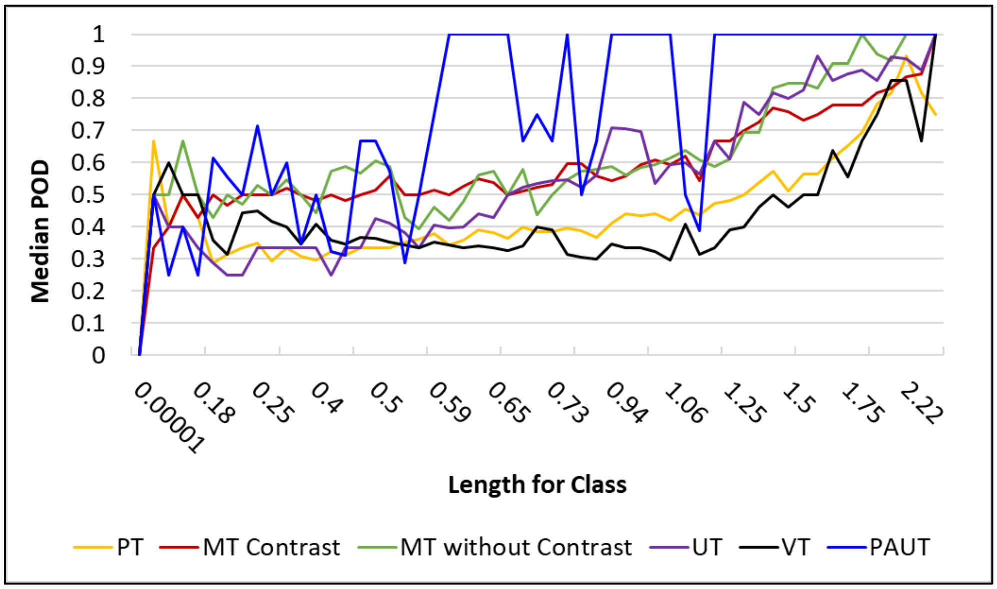


Figure 25. Summary Plot Median POD – BW

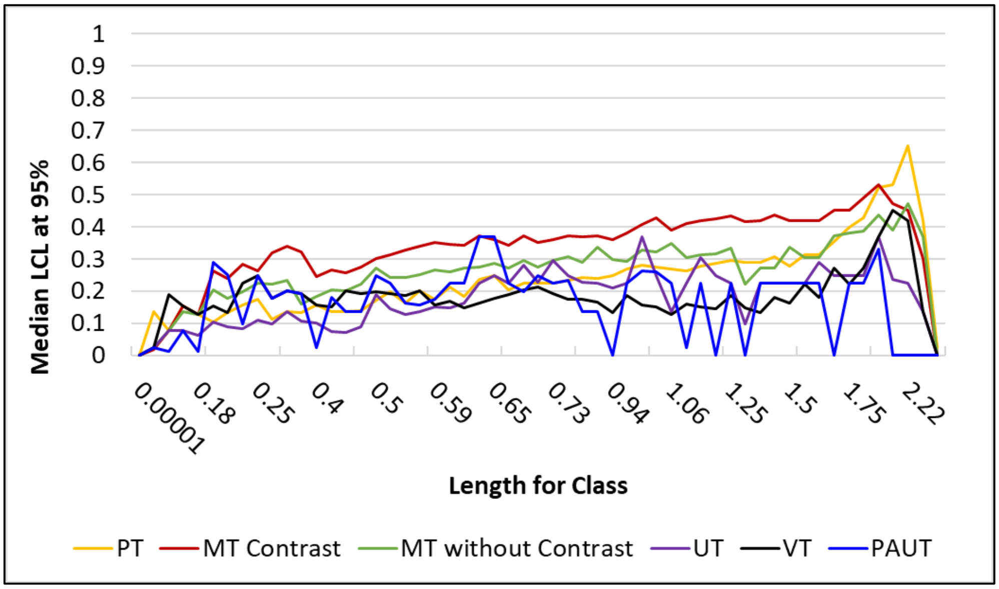


Figure 26. Summary Plot Median LCL at 95 Percent POD – BW

5. Conclusions

TTCI, under the sponsorship of FRA and with the participation from the railroad tank car industry, performed a statistical assessment to evaluate capabilities of CFR-approved NDE methods on the railroad tank cars FW and BW panels. Researchers analyzed a total of 197 POD railroad tank car NDE datasets (both for the BW and FW) from 70 NDE operators obtained from 1996 to 2016 while using three different statistical approaches.

This analysis allowed researchers to make the following conclusions:

TTCI used a traditional statistics methodology, based on the relative frequency of hits and misses, to calculate the POH for each inspection method. From the results, it was observed that this methodology, although it is very simple to implement, is limited because it assumes that each crack size has the same probability of occurrence, since the POH is calculated by dividing the total number of hits in a given a crack length range by the total number of observations in that interval. In addition, the sample size is another limitation of this method. One of the requirements to obtain the probability is to have a large sample size to obtain a more accurate POH for each crack length range. Finally, the results of this method do not meet sufficient statistical confidence levels.

- The MLE-LOGIT 90/95 POD criteria may not be adequate or applicable for all scenarios for NDE system validation for fracture critical railroad tank car fusion welded components. The MLE-based approach had some drawbacks, which included the algorithm convergence issues as well as the results from this method showed that the POD increases as the crack length increases. From the preliminary results observed in this research, there were cases where the opposite case occurred where operators were missing large crack sizes.
- The 90/95 POD metric for NDE inspection capability was originally derived from Mil-HDBK-5H and is now widely accepted in the NASA, Air Force, and other industry applications. Although, this approach established many of the requirements in current tank car inspection specifications and identified as a possible goal for use in railroad tank car NDE inspections during the initial discussions of the HM-201 rulemaking, there is still an ongoing debate within the tank car industry if this is a correct metric for all components of the fusion welded tank cars, and future research must address this.
- DOEPOD approach was more conservative than the other two approaches. DOEPOD findings provided an indication of the non-monotonically increasing POD behavior and demonstrated that the POD value does not increase as the crack length increases. Since the MLE-LOGIT curve fit method for estimating the POD failed with a divergence curve or warning and the curve fit shown could not be used, the DOEPOD recommendation was to validate the POD. Also, in many cases after the validation of the POD at a specific flaw size, when the flaw size was changed, the inspection procedure or inspection system was inadequate, and it is possible that the training of the operator was not adequate or the inspection system may not be applicable to meet the inspection requirement. DOEPOD recommendations are that the inspection system may not be appropriate for meeting inspection criteria, or there is a need to expand the current range of X_L by adding 29 new samples with class lengths of 2XL or greater.

In summary, based on the results obtained, the research team demonstrated that the NDE methods and procedure used for BW failed to reach 90/95 POD. An evaluation of the FW data showed mixed results, but only the MT method reached 90/95 POD. VT and PT methods are the ones that have the lowest POD for all crack lengths for both welds. Both BW and FW datasets observed excessive False Calls, which suggests a lack of operator experience with fatigue cracks or using NDE methods. Researchers strongly recommend a refresher training course especially if an inspector is new to a company.

All inspectors should be trained and capable to detect both manufacturing defects and fatigue related cracks, not just one or the other. The NDT research team's observations during these inspections, and evaluations of results using DOEPOD, detected significant variations in training procedures for inspectors among the industry. Researchers recommend that adequate specialized training on all CFR-approved NDE methods is necessary for personnel inspecting tank cars, regardless of the manufacturing or revenue service location.

Also, the DOEPOD analysis determined that a more thorough calibration step should be followed and written into the procedure for each company. This includes performing a calibration using the type of material being inspected instead of calibration blocks.

Also, quantification of baseline capabilities (POD) is essential to standardization of applicable NDE procedure/methods. Researchers observed in the POD studies used different instruments with different procedures, which may have also contributed to the variations of POD results. The research team recommends the development of standardized or generalized NDE procedure and its validation for use in the industry.

6. References

Advisory Group for Aerospace Research and Development Group. (1993). *A Recommended Methodology for Quantifying NDE/NDI Based on Aircraft Engine Experience.* AGARD-LS-190: North Atlantic Treaty Organization.

Archuleta, M., Poudel, A., Rummel, W. D., & Gonzalez, F. (2016). *Probability of Detection Evaluation Results for Railroad Tank Car Nondestructive Testing*. Technical Report No. DOT/FRA/ORD-16/35, Washington, DC: U.S. Department of Transportation, Federal Railroad Administration.

Association of American Railroads. (2014). *AAR Manual of Standards and Recommended Practices Section C-III Specifications for Tank Cars, M-1002, Appendix T Nondestructive Examination.* Washington, DC: AAR.

Berens, A. P., & Hovey, P. W. (1983). *Evaluating POD/CL Characterizations of NDE Reliability*. Dayton, OH: University of Dayton Research Institute.

Code of Federal Regulations. (2003). *Title 49 CFR Section 179.7, Quality assurance program*. Washington, DC: govregs.

Code of Federal Regulations. (2012). *Title 49 CFR Section 180.509(e), Requirements for inspection and test of specification tank cars, paragraph (e) "Structural integrity inspections tests*. Washington, DC: govregs.

Garcia, G. (2002). *Railroad Tank Car Nondestructive Methods Evaluation*. Technical Report No. DOT/FRA/ORD-01/04, Washington, DC: U.S. Department of Transportation, Federal Railroad Administration.

Garcia, G., Rummel, W. D., & Gonzalez, F. (2016). *Quantitative Nondestructive Testing of Railroad Tank Cars Using the Probability of Detection Evaluation Approach*. Technical Report No. DOT/FRA/ORD-09/10, Washington, DC: U.S. Department of Transportation, Federal Railroad Administration.

Garcia, G., Welander, L., Rummel, W. D., & Gonzalez, F. (2016). *Probability of Detection Evaluation Results for Railroad Tank Cars*. Technical Report No. DOT/FRA/ORD-16/13, Washington, DC: U.S. Department of Transportation, Federal Railroad Administration.

Generazio, E. R. (2009, June). Design of Experiments for Validating Probability of Detection Capability of NDT Systems and for Qualification of Inspectors. *Materials Evaluation, 67*(6), 730-738.

Generazio, E. R. (2011). *Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications*. Report No. NASA/TP-2011-217176. Hampton, VA: National Aeronautics Space Administration.

Generazio, E. R. (2014). *Interrelationships Between Receiver/Relative Operating Characteristics Display, Binomial, Logit, and Bayes' Rule Probability of Detection Methodologies*. Report No. NASA/TM-2014-218183, Hampton, VA: National Aeronautics and Space Administration.

Generazio, E. R. (2015). *Directed Design of Experiments for Validating Probability of Detection Capability of NDE Systems (DOEPOD).* Report No. NASA/TM-2015-218696, Hampton, VA: National Aeronautics Space Administration.

Lewis, W. H., Dodd, B. D., Sproat, W. H., & Hamilton, J. M. (1978). *Reliability of Nondestructive Inspections - Final Report*. Report No. SA-ALC/MEE 76-6-38-1, San Antonio, TX: U.S. Air Force.

National Institute of Standards & Technology. (2012). *NIST/SEMATECH e-Handbook of Statistical Methods*.

National Transportation Safety Board. (1992). *Safety Recommendations R-92-021, R-92-022, R-92-023, R-92-024*. Washington, DC: ntsb.gov.

Pettit, D. E., & Hoeppner, D. W. (1972). *Fatigue Flaw Growth and NDT Evaluation for Preventing Through Cracks iin Spacecraft Tankage Structures*. Report No. NAS 9-11722 LR25387, Houston, TX: National Aeronautics Space Administration.

Rummel, W. D. (1997). *Quantitative Nondestructive Evaluation Capabilities (POD) in Relation to HM-201 Rulemaking.* D&W Enterprises LTD.

Rummel, W. D. (April 16-20, 2010). Nondestructive Inspection Reliability - History, Status and Future Path. *Proceedings for 18th World Conference on Nondestructive Testing.* Durban, South Africa: D&W Enterprises, LTD.

Rummel, W. D., Rathke, R. A., Todd, P. H., & Mullert, S. J. (1975). The Detection of Tightly Closed Flaws by Nondestructive Testing (NDT) Methods. Report No. NASA-CR-144639, Houston, TX: National Aeronautics Space Administration.

Rummel, W. D., Rathke, R. A., Todd, P. H., Tedrow, T. L, & Mullen, S. J. (1976). *Detection of Tightly Closed Flaws by Nondestructive Testing Methods in Steel and Titanium.* Report No. NASA-CR-151098, Houston, TX: National Aeronautics Space Administration.

Rummel, W. D., Todd, P. H., Frecska, S. A., & Rathke, R. A. (1974). The Detection of Fatigue Cracks by Nondestructive Testing Methods. (A. S. Testing, Ed.) *Materials Evaluation, 32*(10), 205–212.

U.S. Department of Defense. (1998). *Metallic Materials and Elements for Aerospace Vehicle Structures*. Report No. MIL-HDBK 5: U.S. Air Force.

U.S. Department of Defense. (2004). *MIL-HDBK-1823A Nondestructive Evaluation System Reliability Assessment, Appendix G - Statistical Analysis of NDE Data*. Wright-Patterson AFB: DOE.

7. Appendix

Federal Railroad Administration. 2021. Analysis of Historical Non-Destructive Evaluation Probability of Detection Data for Railroad Tank Cars: Appendices A Through I. Report No. DOT/FRA/ORD-21/14. Washington, DC: U.S. Department of Transportation.

Abbreviations and Acronyms

ACRONYMS	EXPLANATION
AE	Acoustic Emissions
AGARD	Advisory Group for Aerospace Research and Development Group
AAR	Association of American Railroads
BW	Butt Weld
X_{POD}	Class length at which the lower confidence limit (value) is 0.90 or greater (90/95 POD) @ 95% confidence
$X_{BEST\ LCL}$	Class length (flaw size) exhibiting the maximum LCL
X_{POH}	Class length where there are no misses above this class length, and POH = 1 above this class length
CFR	Code of Federal Regulations
DOEPOD	Design of Experiments Probability of Detection
EDM	Electrical Discharge Machine
POH	Estimate of Probability of Hit (Number of Hits in Class Length/Total Number of Trials in Class Length)
FCP	False Call Percentage
FRA	Federal Railroad Administration
FW	Fillet Weld
HMR	Hazardous Materials Regulations
X_L	Largest Class Length in Entire Dataset
LOGIT	Logic Regression Method
LCL	Lower Confidence Level (value) of POH @ 95% confidence
MT	Magnetic Particle Testing
MLE	Maximum Likelihood Estimation
NASA	National Aeronautics and Space Association
NTSB	National Transportation Safety Board
NDE	Nondestructive Evaluation
NDT	Nondestructive Testing
PT	Penetrant Testing
PAUT	Phased Array Ultrasonic Testing
POD	Probability of Detection (the true POD obtained if an infinite number of samples are used)

ACRONYMS	EXPLANATION
PROBIT	Probit Progressive Method
QAP	Quality Assurance Program
RT	Radiographic Testing
TTC	Transportation Technology Center (the site)
TTCI	Transportation Technology Center, Inc. (the company)
MLE-LOGIT	Two-parameter Logit Model
UCB	Upper Confidence Bound
DOT	U.S. Department of Transportation
UT	Ultrasonic Testing
VT	Visual Testing