ERDC/EL TR-24-12

Environmental Laboratory

U.S. ARMY

US Army Corps of Engineers®

ERDC

Navigation Systems Research Program (NAVSYS)

A Study of Phased-Array Ultrasonic Testing (PAUT) for Detecting, Sizing, and Characterizing Flaws in the Welds of Existing Hydraulic Steel Structures (HSS)

Martin T. Schultz, Leslie E. Campbell, Ramsay D. Bell, and Phillip W. Sauser

June 2024

The US Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdc.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at https://erdclibrary.on.worldcat.org/discovery.

ERDC/EL TR-24-12
June 2024

A Study of Phased-Array Ultrasonic Testing (PAUT) for Detecting, Sizing, and Characterizing Flaws in the Welds of Existing Hydraulic Steel Structures (HSS)

Martin T. Schultz

US Army Engineer Research and Development Center (ERDC)
Environmental Laboratory (EL)
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Leslie E. Campbell

US Army Corps of Engineers (USACE)
New Orleans District
7400 Leake Avenue
New Orleans, LA 70118

Ramsay D. Bell

US Army Corps of Engineers (USACE)
Welding and Metallurgy Technical Center of Expertise
333 SW 1st Avenue
Portland, OR 97204

Phillip W. Sauser

US Army Corps of Engineers (USACE)
Northwestern Division
1201 NE Lloyd Ste 400
Portland, OR 97232

Final Technical Report (TR)

Prepared for US Army Engineer Research and Development Center
Coastal and Hydraulics Laboratory
3909 Halls Ferry Road, Vicksburg, MS 39180-6199

Under Funding Account Code U4388268, AMSCO Code 031391

Abstract

Hydraulic steel structures (HSS) are components of navigation, flood control, and hydropower projects that control or regulate the flow of water. Damage accumulates in HSS as they are operated over time, and they must be inspected periodically. This is often accomplished using nondestructive testing (NDT) techniques. If damage is detected, the structure's fitness for continued service must be evaluated, which requires information on the location and size of discontinuities. This information can be obtained using ultrasonic testing (UT) techniques. However, there is limited information on the reliability of UT techniques with respect to detecting, sizing, and characterizing flaws in HSS. This study addresses this gap. Round-robin experiments were carried out using phased-array ultrasonic testing (PAUT) to scan weld specimens representing a variety of HSS geometries. The results of the round-robin experiments were analyzed to estimate the probability of detection (POD) and to assess the influence of factors potentially affecting POD. Uncertainty in estimates of flaw length and height were described, and partial safety factors were derived for use in fitness-for-service analyses. These results demonstrate the importance of the technician as a factor influencing the reliability of NDT techniques applied to HSS.

Contents

Abstract ii

Figures and Tables vi

Preface xi

1 Introduction 1

1.1 Background 1
1.1.1 Nondestructive Testing (NDT) Techniques 3
1.1.2 Certification of NDT Technicians 9
1.1.3 Round-Robin Testing and the Reliability of Ultrasonic Testing (UT) 10
1.2 Objectives 17
1.3 Approach 17

2 The Design and Manufacture of Weld Specimens 21

2.1 The Design of US Army Engineer Research and Development Center (ERDC) Weld Specimens 21
2.2 Verification of the Location and Size of Flaws in Weld Specimens 30

3 Develop and Validate Ultrasonic NDT Procedures 33

3.1 Expert Workshop on UT 33
3.2 Testing, Refinement, and Validation of Initial Procedures 33
3.3 The Results of Procedure Validation 41
3.3.1 Flaw Detection 42
3.3.2 Flaw Sizing 47
3.3.2.1 Flaw Length 47
3.3.2.2 Flaw Height 51
3.3.3 Flaw Characterization 54

4 Round-Robin Testing 57

4.1 Performance Qualification 57
4.2 Round-Robin Testing 61
4.3 Round-Robin Testing Results 65
4.3.1 Diagnostic Test Evaluation 67
4.3.2 Technician Performance Metrics 68
4.3.3 Correlations among Performance Metrics 69
4.3.4 Logistic Regression Models for Probability of Detection (POD) 73
4.3.4.1 Planar Flaws 77
4.3.4.2 Volumetric Flaws 79
4.3.4.3 Laminar Flaws 81
4.4 POD Model Sensitivity 83
4.5 POD Estimates by Flaw 89
4.6 Reliability of Reported Indications 92

4.7 Uncertainty in Estimates of Flaw Length and Height 96
4.7.1 Uncertainty by Flaw and Joint Category 102
4.7.2 Uncertainty by Flaw Subcategory 105
4.7.3 Uncertainty by Technician 108
4.7.4 Partial Safety Factors 111
4.8 Flaw Characterization 114
4.8.1 Three-Level Classification System 115
4.8.2 Six-Level Classification System 116
4.8.3 Ten-Level Classification System 117
4.8.4 Variability in Accuracy and Reliability among NDT Technicians 118

5 Application of Partial Safety Factors to Fitness-for-Service (FFS) Examples 121
5.1 Failure Assessment Diagram (FAD) 121
5.2 Partial Safety Factors 123
5.3 Fracture Evaluation Example 1: Flat Plate with Surface Crack 123
5.3.1 Example Problem Description 124
5.3.2 Calculate Load Ratio, L_r 125
5.3.3 Calculate Fracture Ratio, K_r 126
5.3.4 Develop FAD 126
5.4 Fracture Evaluation Example 2: Flat Plate with Embedded Crack 128
5.4.1 Example Problem Description 129
5.4.2 Calculate Load Ratio, L_r 130
5.4.3 Calculate Fracture Ratio, K_r 130
5.4.4 Develop FAD 131

6 Discussion and Recommendations 133
6.1 NDT Techniques and Procedures 133
6.2 Finding NDT Technicians 134
6.3 Technician Training and Prequalification 135
6.4 POD 138
6.5 Reliability of Indications 139
6.6 Uncertainty in Flaw Size Estimation 140
6.7 Flaw Characterization 142
6.8 Verification Sampling 143
6.9 Summary of Recommendations 143

References 145

Appendix A: Examples of Representative Hydraulic Steel Structure (HSS) Joint Geometries 151

Appendix B: Fabrication Drawings and Photographs of US Army Engineer Research and Development Center (ERDC) Specimens 161

Appendix C: As-Built Drawings of American Welding Society (AWS) Specimens 186

Appendix D: Minutes of the Expert Workshop on Nondestructive Testing (NDT) 195

Appendix E: NDT Procedures 203

Appendix F: Complete Calculations for Fitness-for-Service (FFS) Examples 218

Abbreviations 244

Report Documentation Page (SF 298) 248

Figures and Tables

Figures

1. Flaw detected in the first leg of a sound path using pulse-echo (PE) ultrasonic testing (UT). 4
2. Flaw detected in the second leg of the sound path using PE UT. 5
3. Flaw detection using time-of-flight diffraction (TOFD). 5
4. Signal intensity within a sound beam. 6
5. Phased-array ultrasonic testing (PAUT) using multiple probes and varying signal angles to achieve complete volumetric coverage. 7
6. Comparison of (*left*) PAUT and (*right*) total focus method (TFM) / full matrix capture (FMC) scans. 8
7. Representative welded steel specimens with faces and coordinates labeled. 24
8. Conventions used in labeling faces and axes of three different joint types. 25
9. Distribution of flaws by flaw subcategory. 26
10. Distribution of flaws by length. 28
11. Distribution of flaws by height. 28
12. Illustration of flaw coordinates. 29
13. Estimated and actual flaw length from verification testing. 32
14. Estimated and actual flaw height from verification testing. 32
15. Level III technicians during inspection trials to test and validate procedures. 35
16. Common issue with the coordinate system on US Army Engineer Research and Development Center (ERDC) 005. 37
17. Flowchart illustrating the system for distinguishing between hits and false calls. 38
18. Plan view of a specimen showing the location of flaws (*solid lines*) and reported indications (*dashed lines*) along the *y*-axis. 39
19. Section view of a specimen showing the location of flaws (*solid lines*) and reported indications (*dashed lines*) along the *z*-axis. Flaw 2 is a transverse flaw with a height of approximately 0.2 in. and a length of approximately 1 in. Opposite corners of the flaw are indicated in the figure. 40
20. Comparison of estimated and actual flaw length using PAUT. 49
21. Comparison of estimated and actual flaw length using TFM/FMC. 50
22. Comparison of estimated and actual flaw height using PAUT. 53
23. Comparison of estimated and actual flaw height using TFM/FMC. 53
24. Level II technicians during round-robin inspections. 64
25. Number of hits (*black*), misses (*stippled*), and false calls (*gray*) by technician for PAUT scans (ERDC and American Welding Society [AWS] specimens). 72
26. Fraction of PAUT scans during which the flaws were missed. 73
27. Probability of detection (POD) for the nominal planar flaw when varying flaw height and flaw aspect. 87
28. POD for the nominal volumetric flaw when varying plate thickness and flaw aspect. 88

29. POD for the nominal laminar flaw when varying technician true positive rate (TPR) and lamination area. 89
30. Estimated POD and observed round-robin detection rate for each flaw. 92
31. Reliability of reported indications as a function of reported flaw length and aspect. 96
32. Actual and estimated flaw length and height. 97
33. Error in estimates of flaw length and height. 98
34. Root-mean-square error (RMSE) by actual flaw length and height. 99
35. Ratios of estimated to actual flaw length and height. 100
36. Cumulative distribution functions describing uncertainty in length and height estimates. 101
37. Quantile–quantile (Q-Q) plots for lognormal distributions fit to the ratio of estimated to actual length and height. 102
38. Uncertainty in the ratio of estimated to actual length and height by flaw category and joint type. 103
39. Q-Q plots for lognormal distributions fit to the ratio of estimated to actual length and height by flaw category. 104
40. Q-Q plots for lognormal distributions fit to the ratio of estimated to actual length and height by joint type. 105
41. Cumulative distribution functions showing uncertainty in the ratio of estimated to actual length and height by flaw subcategory. 106
42. Q-Q plots showing the fit of lognormal distributions to the ratio of estimated-to-actual length by flaw subcategory. 107
43. Q-Q plots showing the fit of lognormal distributions to the ratio of estimated to actual height by flaw subcategory. 108
44. Cumulative distribution functions showing uncertainty in estimates of flaw length and height by technician identifier. 110
45. Partial safety factors by technician. The *black square* is for the Level III technician, and the *black circles* are for the Level II technicians. 113
46. Sensitivity (SEN) and positive predictive value (PPV) for each flaw category in the 10-level classification system. 118
47. Variability in flaw classification accuracy across nondestructive testing (NDT) technicians for each of the three classification systems. 120
48. Schematic representation of an Option 1 failure assessment diagram (FAD). 122
49. Schematic representation of an Option 2 FAD. 122
50. Plate and flaw geometry for a surface flaw. (Image adapted from BS-7910:2013+A1:2015 [BSI 2015], Figure M.3.) 125
51. FAD for three surface cracks in a carbon steel plate. Assessment points based on the reported flaw size (without partial safety factors) are *circles*, and assessment points for the analyzed flaw size (with partial safety factors) are *squares*. 128
52. Plate and flaw geometry for an embedded flaw. (Image adapted from BS-7910:2013+A1:2015 [BSI 2015], Figure M.8). 128

53. FAD for three embedded cracks in a carbon steel plate. Assessment points based on the reported flaw size (without partial safety factors) are *circles*, and the assessment points for the analyzed flaw size (with partial safety factors) are *squares*. 132
A-1. Examples of Joint A (US Army Engineer Research and Development Center [ERDC] 001) and Joint F (ERDC 009). 152
A-2. Example of Joint A (ERDC 001). 153
A-3. Example of Joint B (ERDC 002). 154
A-4. Examples of Joint C (ERDC 004) and Joint H (ERDC 012). 155
A-5. Examples of Joint D (ERDC 005) and Joint B (ERDC 003). 156
A-6. Example of Joint F (ERDC 008). 157
A-7. Examples of Joint E (ERDC 007) and Joint G (ERDC 010). 158
A-8. Example of Joint E (ERDC 007). 159
A-9. Examples of Joint H (ERDC 011). 160
B-1. Fabrication drawing for ERDC 001. 162
B-2. Photograph of ERDC 001. 163
B-3. Fabrication drawing of ERDC 002. 164
B-4. Photograph of ERDC 002. 165
B-5. Fabrication drawing for ERDC 003. 166
B-6. Photograph of ERDC 003. 167
B-7. Fabrication drawing for ERDC 004. 168
B-8. Photograph of ERDC 004. 169
B-9. Fabrication drawing of ERDC 005. 170
B-10. Photograph of ERDC 005 171
B-11. Fabrication drawing of ERDC 006. 172
B-12. Photograph of ERDC 006 173
B-13. Fabrication drawing of ERDC 007. 174
B-14. Photograph of ERDC 007 175
B-15. Fabrication drawing for ERDC 008 176
B-16. Photograph of ERDC 008 177
B-17. Fabrication drawing of ERDC 009. 178
B-18. Photograph of ERDC 009 179
B-19. Fabrication drawing of ERDC 010. 180
B-20. Photograph of ERDC 010 181
B-21. Fabrication drawing of ERDC 011. 182
B-22. Photograph of ERDC 011 183
B-23. Fabrication drawing of ERDC 012. 184
B-24. Photograph of ERDC 012 185

Tables

1. Examples of round-robin testing experiments in the literature. 14
2. Specimen matrix for custom-built and off-the-shelf test specimens. 23
3. Flaw categories, subcategories, definitions (AWS 2020b), and dimensions in US Army Engineer Research and Development Center (ERDC) and American Welding Society (AWS) specimens. 27
4. Roster of participating Level III nondestructive testing (NDT) technicians. 34
5. Summary of Level III inspection trials showing in what phase of testing (A, B, or C) each specimen was scanned by each technician. 36
6. Evaluation of indication 1. 40
7. Evaluation of indication 2. 41
8. Evaluation of indication 3. 41
9. Summary of verification test results. 42
10. Detection rates by joint type. 43
11. False calls by joint type. 43
12. Hit-to-call ratio by joint type. 43
13. Detection rate (total number of possible hits in parentheses) by specimen. 43
14. False calls by specimen. 44
15. Detection rate (total number of possible hits in parentheses) by flaw type. 45
16. Detection rate (total number of possible hits in parentheses) by technician. 45
17. False calls by technician. 45
18. Hit-to-call ratio by technician. 46
19. Detection rate (total number of possible hits in parentheses) by flaw length. 46
20. Error analysis for flaw length estimates by NDT technique. 48
21. Error analysis for flaw length measurements by joint type. 51
22. Error analysis for flaw height measurements. 52
23. Accuracy in flaw characterization. 54
24. Percent accuracy of flaw characterization by NDT technique and testing phase. 55
25. Percent accuracy in flaw type characterization using PAUT. 55
26. Percent accuracy in flaw type characterization using TFM/FMC. 55
27. Summary of participation in prequalification sessions. 59
28. Performance qualification results. 60
29. Performance qualification results. 61
30. Number and affiliation of Level II NDT technicians by round-robin testing week. 62
31. Equipment used by Level II technicians during round-robin testing. 63
32. Round-robin test results for ERDC specimens. 65
33. Round-robin test results for AWS D1.5 Bridge Kit specimens. 66
34. Round-robin test results, ERDC and AWS specimens combined, using PAUT. 67
35. Confusion matrix example. 68

36. Performance qualification and round-robin test performance metrics. 69
37. Pearson correlation coefficients for performance metrics calculated from round-robin scans of AWS and ERDC specimens. 70
38. Pearson correlation coefficients for performance metrics by specimen class. 71
39. Independent variables used in logistic regression by flaw category. 74
40. Levels for flaw category and subcategory. 75
41. Descriptive statistics for continuous independent variables used in logistic regression by flaw category 76
42. Logistic regression results for probability of detection (POD) of planar flaws. 78
43. Logistic regression results for POD of volumetric flaws. 80
44. Logistic regression results for POD of laminar flaws. 82
45. Sensitivity of detection probability to significant independent variables. 84
46. Estimates of the POD from logistic regression models for all 68 flaws scanned during round-robin testing 90
47. Logistic regression results for the reliability of an indication. 94
48. Descriptive statistics for continuous independent variables used to estimate the nominal reliability of reported indications 95
49. Estimates of the probability that a reported indication is an actual flaw 95
50. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height by flaw and joint type and for all flaws 103
51. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height by flaw category 106
52. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height. 109
53. Partial safety factors for flaw length and height by flaw category and joint type. 112
54. Partial safety factors by technician identification (ID). 113
55. Reported and actual flaw characterization for the classification system with three flaw categories 115
56. Reported and actual flaw characterization for the classification system with six flaw categories 116
57. Reported and actual flaw characterization for the classification system with 10 levels 117
58. Fraction of flaws detected and characterized accurately using the 10-, 6-, and 3-level classification systems 119
59. Summary of results for three surface cracks in a carbon steel plate. 127
60. Summary of results for three embedded cracks in a carbon steel plate. 132
A-1. Descriptions and examples of representative joints used as the basis for US Army Engineer Research and Development Center (ERDC) specimens 151

Preface

This study was conducted for the US Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory (ERDC-CHL), under the Navigation Systems Research and Development Program, Funding Account Code U4388268, AMSCO Code 031391. The technical monitor was Ms. Morgan Johnston.

The work was performed by the Environmental Risk Assessment Branch of the Environmental Processes and Engineering Division, ERDC Environmental Laboratory (EL). Mr. James Lindsay was chief, Environmental Risk Assessment Branch; Mr. Warren P. Lorentz was chief, Environmental Processes and Engineering Division; and Mr. Charles E. Wiggins was the technical director for Navigation. The deputy director of ERDC-EL was Dr. Brandon Lafferty, and the director was Dr. Edmond J. Russo.

The authors wish to acknowledge numerous individuals who assisted with this study. Dr. Matthew D. Smith (ERDC, Geotechnical and Structures Laboratory [GSL]) assisted with organizing and conceptualizing the project in its early stages. Members of the USACE Structural Community of Practice provided guidance on selecting hydraulic steel structure (HSS) joint geometries to be represented in weld specimens. Mr. Travis Filmore (ERDC-CHL) developed an experimental design for flaw and specimen characteristics used in round-robin experiments. Mr. James Kinnebrew (ERDC, Information Technology Laboratory [ITL]) and Mr. Jason Ray (ERDC-ITL) assisted with data recording during the validation of nondestructive testing (NDT) procedures. Mr. Alan Katzenmeyer assisted with the transportation of equipment and specimens for round-robin testing. Dr. Francesco Russo (Michael Baker International [MBI]) led an expert workshop to discuss the development of NDT procedures for HSS. That expert panel included Dr. Robert Connor (Purdue University), Dr. Karl Frank (University of Texas at Austin), Dr. Hoda Azari (Federal Highway Administration [FHWA]), Mr. Russell Kok (FHWA), Mr. Ronnie Medlock (High Structures Steel, Inc.), Mr. Jordan Wind (Bureau Veritas), Mr. Parrish Furr (Loenbro), and Mr. Ray Momsen (Bureau Veritas). Dr. Francesco Russo and Mr. Eric Frary (MBI) recruited and subcontracted NDT technicians to assist with developing and validating NDT procedures and implementing the round-robin experiments. Bureau Veritas, Loenbro, Applied Technical Services, MISTRAS group, and TechKnowServ supplied NDT

technicians for the validation of NDT procedures or round-robin testing. Dr. Guillermo Riveros (ERDC-ITL), Dr. Curtis Schroeder (Wiss, Janney, Elstner Associates, Inc.), Mr. Travis Adams (USACE Welding and Metallurgy Technical Center of Expertise), and Mr. Nicholas Shrawder (US Navy) provided comments on a draft version of this technical report. Mr. Christopher Westbrook, USACE Structural Community of Practice lead, provided comments and suggestions on the revised manuscript.

COL Christian Patterson was the commander of ERDC, and Dr. David W. Pittman was the director.

1 Introduction

1.1 Background

Hydraulic steel structures (HSS) are structures that control or regulate water and are typically part of a larger navigation, hydropower, or flood control project. Typical HSS include lock gates, dam spillway gates, valves, bulkheads and stoplogs, vertical lift gates, components of hydroelectric and pumping plants, and miscellaneous structures such as lock wall accessories, flood protection gates, lifting beams used for installing other HSS, and outlet works gates (USACE 2009). Early construction methods consisted of connecting members with rivets. Welded construction became popular in the 1950s and 1960s. Riveted construction offered the advantage of internal redundancy, where fractures occurred in only the local member (i.e., plate or rivet) and could not propagate to other members. The disadvantage of riveted construction was the infiltration of moisture between components and the resulting damage (e.g., pack rust), increased labor, and the additional material required to make the lap joint connection. Welded construction offered the advantage of efficiency and the ability to seal connected plates, the latter avoiding corrosion and pack rust between members. The disadvantage of welded construction was that damage, particularly cracking, could propagate into all connected members. This problem was exacerbated through a lack of understanding of welding-related defect mechanisms (e.g., cracking), improper welding procedures, poor welding quality, and a lack of understanding of fatigue and fracture, the primary causes of cracking.

Damage accumulates in HSS as they are operated over time, and periodically, it is necessary to inspect these structures and, if flaws are detected, to evaluate their fitness for service (FFS). During an inspection, a concerted effort is made to identify discontinuities that may affect the load carrying capacity of the steel members. Of particular concern are discontinuities in fracture-critical members that may cause catastrophic collapse of the structure. The American Welding Society (AWS) defines a discontinuity as an interruption of the typical structure of a material, such as a lack of homogeneity in its mechanical, metallurgical, or physical characteristics (AWS 2020b). Inspection methods include visual and other methods of nondestructive testing (NDT) that provide full volumetric testing. When discontinuities are found in structural members, engineers must apply industry standards such as the AWS Structural Welding Code D1.1 (AWS 2020c) or the AWS Bridge Welding

Code D1.5 (AWS 2020a) to determine its acceptability. If the discontinuity does not satisfy the industry standards for acceptability, then an FFS analysis is required to demonstrate its acceptability or it must be repaired. FFS analysis requires a strength, fracture, and fatigue analysis and incorporates information about the material properties; the location, size, and character of the discontinuity; and the applied stresses. Agency guidance on inspection, evaluation, and repair of HSS is provided in Engineering Manual (EM) 1110-2-6054 (USACE 2001). This manual also includes specific procedures for FFS analysis and acceptance criteria. These methods are based largely on a guidance document published by the British Standards Institute (BSI 2005) and referred to as BS 7910:2005. Although subsequent editions of BS 7910, including BS 7910:2013+A1:2015 (BSI 2015) and BS 7910:2019 (BSI 2019) provide improved methodologies, BS 7910:2005 is still valid.

The US Army Corps of Engineers (USACE) is currently revising EM 1110-2-6054 (USACE 2001) and updating its FFS analysis methodology. Since the guidance was last updated, there have been improvements in NDT techniques and equipment and in FFS methods. This study aims to inform the revision of that guidance by providing information on the ability to detect, size, and characterize discontinuities in HSS using selected ultrasonic testing (UT) techniques. UT works by propagating high frequency sound waves through a solid object. If discontinuities are present, the sound waves are interrupted and reflected back to the instrument. The instrument displays information provided by the reflected sound waves on a screen. The information and any irregularities are then interpreted by an NDT technician. The effectiveness of UT may be limited in some materials. For example, UT resolution and sensitivity are often sacrificed when testing coarse grain materials. The effectiveness of UT may also be limited by the geometry of a structure. Complex geometry may cause sound waves to reflect in ways that are mistaken for flaws and may limit physical access to the weld. UT techniques have been used effectively in many industries. However, some of the more advanced UT techniques have not been widely adopted within the structural industry. For example, these include phased-array ultrasonic testing (PAUT), time-of-flight diffraction (TOFD), and total focus method (TFM) / full matrix capture (FMC). These have been widely used in the oil and gas, aerospace, railway, and power-generation industries, but they are still an emerging technology within the structural NDT industry and with respect to HSS, specifically.

The terms *flaw*, *defect*, and *discontinuity* are commonly used as synonyms within the NDT industry. However, there are slight, but important, distinctions between the three words as defined by AWS. AWS D1.5 (AWS 2020a) defines a discontinuity as an interruption in the typical structure of the material, and it defines a defect as a discontinuity that renders a part or product unable to meet minimum acceptance standards. The term flaw is not used or defined in current editions of AWS D1.5, but it was used to describe an undesirable discontinuity in previous editions of AWS D1.5. In this report, flaw is used as defined in ASTM E1316-23 (ASTM 2023) and refers to an imperfection or discontinuity that may be detectable by NDT but is not necessarily rejectable.

1.1.1 Nondestructive Testing (NDT) Techniques

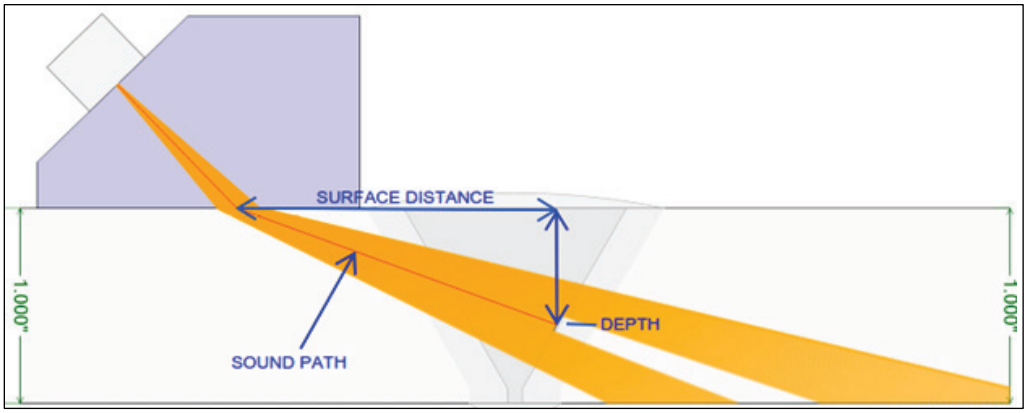
NDT techniques are designed to collect information about imperfections in a solid material without compromising its integrity. Traditional NDT techniques include UT, radiography testing (RT), eddy current testing (ET), magnetic particle testing (MT), and dye penetrating testing (PT) (CNDE, n.d.). UT, RT, and ET are used for detecting and sizing imperfections below the surface. UT uses high-frequency sound waves, RT uses penetrating gamma or X-rays, and ET uses electrical and magnetic fields to accomplish this. MT and PT are used for detecting surface breaking flaws and are generally reliable for identifying the ends of a surface crack or crack length. The preferred NDT techniques in any given situation will depend on a variety of factors, including the type of material, the type of flaw, the orientation and location of the flaw within the thickness of the member, the accessibility of the weld, joint geometry, member thickness, and surface conditions. As these factors will generally not be known prior to an inspection, the type of weld in question is probably the single biggest factor in selecting NDT techniques.

UT and RT are used to size and locate discontinuities through the full weld volume. UT is more commonly used due to the limitations of RT, which include (1) the inability to accurately measure the height or depth of a flaw; (2) the inability to detect cracks oriented perpendicular to the radiation source; (3) the hazards associated with active radiation sources; (4) the difficulty using RT on structural members with complex geometries; and (5) the high expense, particularly in a field environment (AWS 2020a). While AWS D1.5 requires that both UT and RT be used for inspection of fracture-critical butt joints, these limitations suggest that RT should not be used for flaw sizing on in-service HSS. UT is a common NDT technique

used in the evaluation of heavy civil structures, such as pipelines, bridges, and HSS, and is easily procured. UT uses the propagation of ultrasonic waves, typically within the 1 to 10 MHz[1] range in steel, to detect internal discontinuities. UT flaw sizing techniques are classified either as amplitude based (e.g., pulse-echo [PE]) or time based (e.g., TOFD).

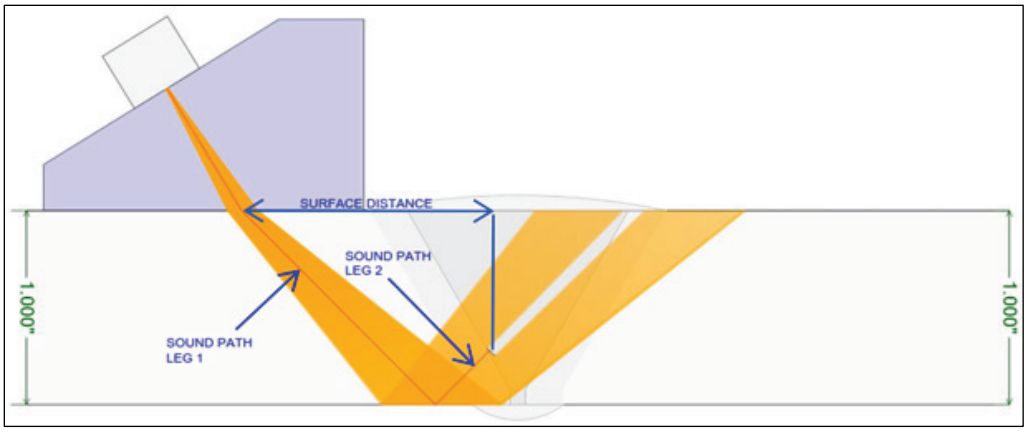
The conventional PE technique transmits a sound wave from a transducer and receives the reflected signal at the same transducer. The received signal is measured in terms of the intensity of the signal or the volume of sound returned and is reported in units of decibels. The location of the detected signal is measured by the amount of time it takes the signal to traverse the sound path from transmit to receive (Figures 1 and 2). In Figure 1, the flaw is located in the first leg (i.e., a direct path from transducer to flaw and back to the transducer). In Figure 2, the flaw is located in the second leg (i.e., refracting off the back surface of the member and reflecting back to the transducer).

Figure 1. Flaw detected in the first leg of a sound path using pulse-echo (PE) ultrasonic testing (UT).



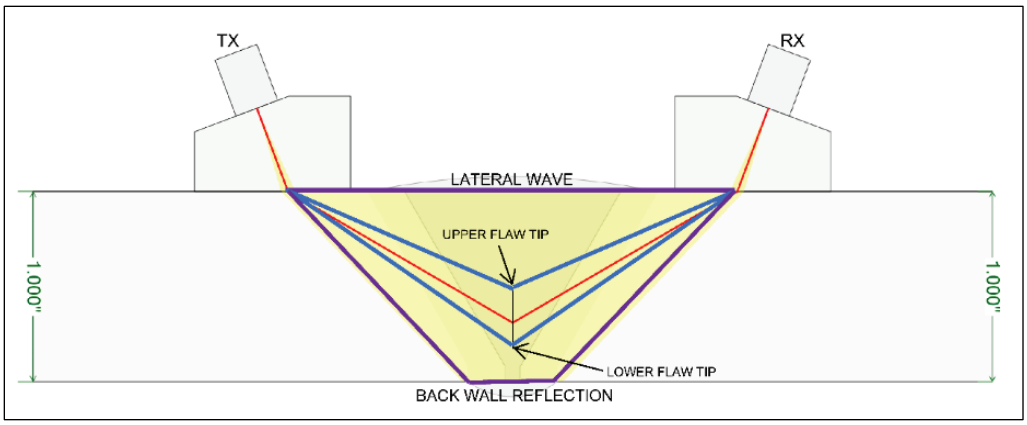
[1] For a full list of the spelled-out forms of the units of measure used in this document, please refer to *US Government Publishing Office Style Manual*, 31st ed. (Washington, DC: US Government Publishing Office, 2016), 248–252, https://www.govinfo.gov/content/pkg/GPO-STYLEMANUAL-2016/pdf/GPO-STYLEMANUAL-2016.pdf.

Figure 2. Flaw detected in the second leg of the sound path using PE UT.



TOFD requires two transducers to be placed on opposite sides of the area to be inspected. One transducer sends the signal, and the other transducer receives the signal. The signal is diffracted as it passes around a flaw (Figure 3). In this example, a signal is diffracted at each end of the flaw and received at the opposite transducer. The location of the flaw tips is discovered by measuring the ultrasonic pulse time of flight and trigonometric measurements.

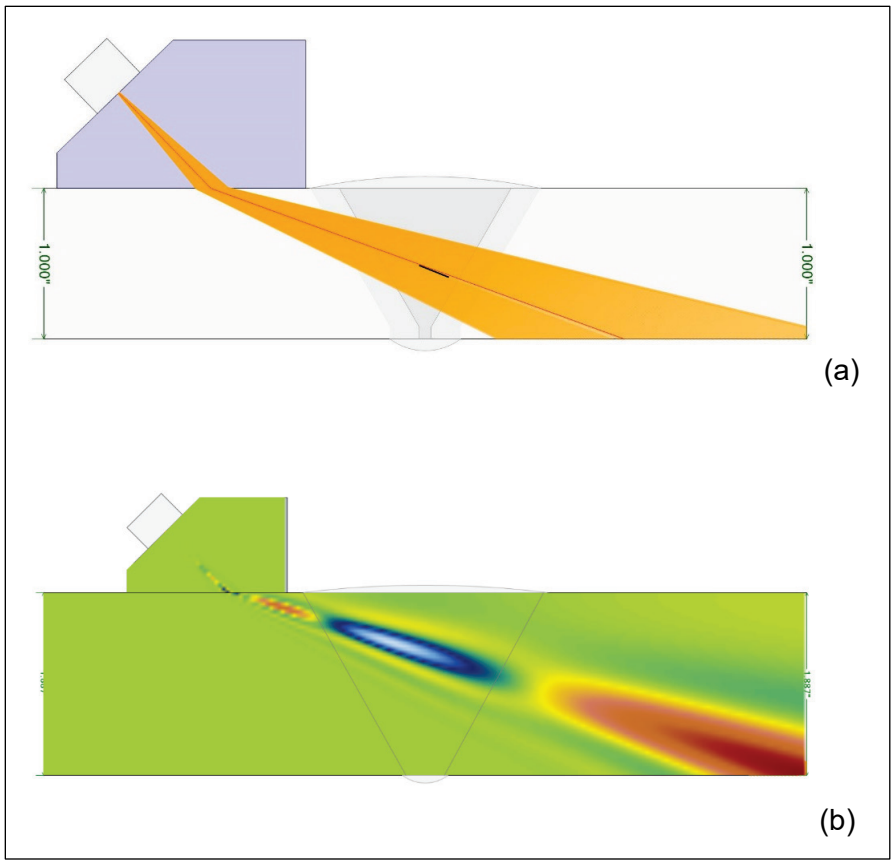
Figure 3. Flaw detection using time-of-flight diffraction (TOFD).



To effectively use UT methods, the limitations must be identified and understood so that they can be accommodated to the extent possible. PE relies on the amplitude of the signal returned, and many of its disadvantages can be attributed to the loss of strength of the returned signal due to less-than-optimal flaw orientation relative to the direction of the signal, beam spread, or attenuation. The strongest signal will be returned from a flaw that is oriented perpendicular to the sound wave, similar to what is shown in Figures 1 and 2. As the orientation deviates from perpendicular, as

shown in Figure 4*a*, the intensity of the returned signal decreases. This is depicted in Figure 4*b*. Within a sound beam, the signal with the greatest intensity is represented by blue, and the signal with least strength is represented by yellow. The more that the orientation deviates from perpendicular to the sound path, the lower the intensity of the signal that is returned to the transducer. It can also be seen from Figure 4*b* that beam size increases (i.e., spreads) the farther away the signal is from the transducer, with an increasing decline in intensity. Attenuation is the decay rate of the signal as it propagates through the material. Some of the intensity of the signal is scattered due to reflection, and some through absorption or a conversion to other forms of energy.

Figure 4. Signal intensity within a sound beam.



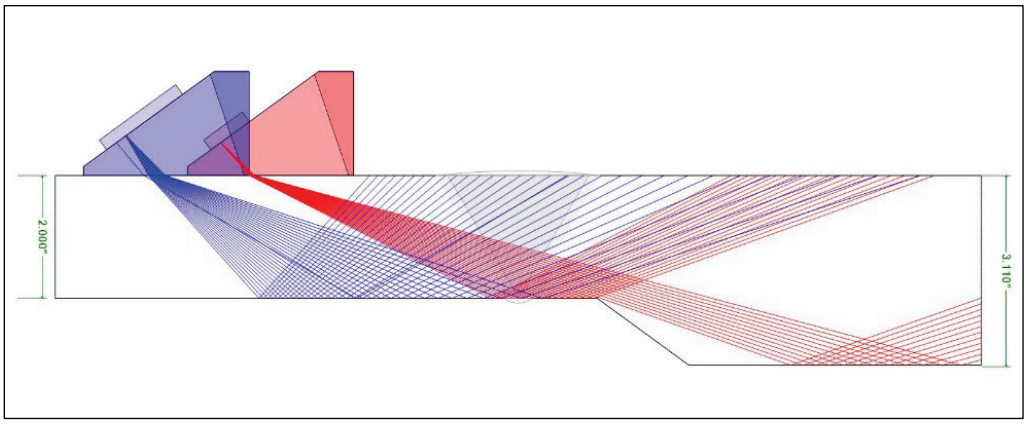
TOFD does not rely solely on the amplitude of the received signal, and the results are not influenced by flaw orientation. Therefore, it does not suffer the same disadvantages as PE. However, TOFD is not free of limitations. The disadvantages of TOFD include the inability to detect flaws (1) near the scanning surfaces and (2) in welds with complex or constrained geometries. The lateral wave at the scanning surface and the backwall echo

interfere with diffraction, and thus, flaws at or near the surface cannot be detected. Constrained and complex geometries may not provide the width of testing surface required to set up dual transducers.

Some of the deficiencies of each method can be mitigated by properly using the equipment and developing procedures specific to a particular scan. However, the best mitigation is combining the two methods to evaluate a member. TOFD complements PE because it does not depend on flaw orientation. PE can detect flaws in regions that are undetectable or inaccessible by TOFD. The deficiencies of each cannot be completely overcome by the other, but together they can improve detection and sizing to produce a more effective and reliable test method.

Technological advancements have mitigated many of the deficiencies in testing. One advancement is PAUT. PAUT utilizes a probe with multiple elements, typically ranging from 16 to 256, arranged in an array. Each transducer sends and receives its own signal, just like the single transducer described previously. Each transducer can be controlled to transmit a unique signal and, when combined with the other signals, increases the detection capabilities relative to the PE technique. When scanning thicker plates or complex geometry, a common approach is to use multiple probes to achieve complete volumetric coverage over a range of 40° to 70° to optimize signal orientation relative to the flaw orientation. Thus, the likelihood of achieving optimal signal orientation is greatly increased. Figure 5 shows an example of a PAUT array with varying signal angles.

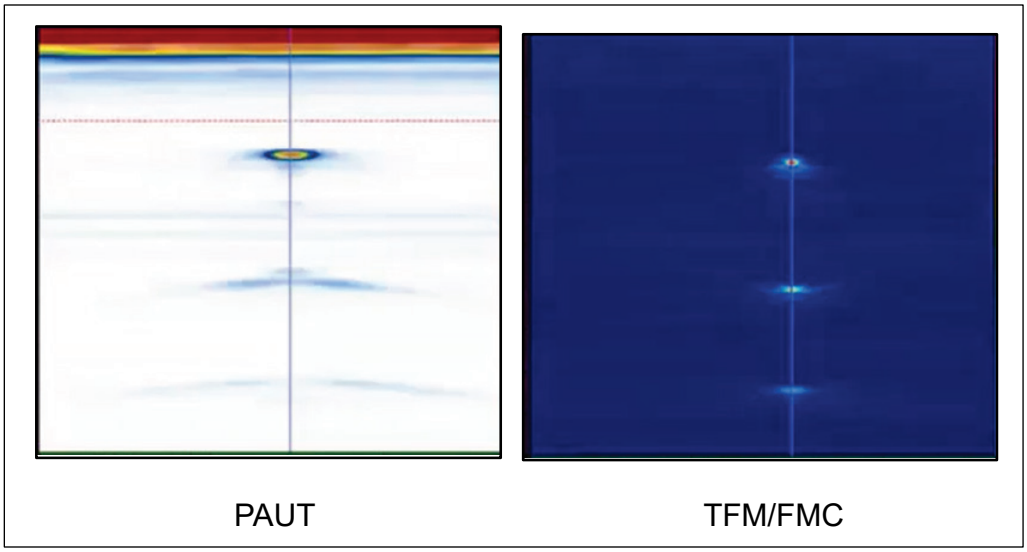
Figure 5. Phased-array ultrasonic testing (PAUT) using multiple probes and varying signal angles to achieve complete volumetric coverage.



TFM/FMC is a more recent advancement in UT that utilizes the same probe as PAUT to capture ultrasound data and software to process that

data. The difference between PAUT and TFM/FMC is that, in TFM/FMC, each element transmits a signal, and all elements receive that signal. The result is n^2 signals received, where n is the number of elements in an array. This results in an exponential increase in data collected in the FMC scan compared to a PAUT scan. TFM is a signal processing algorithm used to evaluate FMC data. The scan cross section is discretized into a predetermined grid. Every data point generated by FMC is summed at each grid point and smoothed to generate the final scan image. The result is a much higher resolution and greater accuracy in sizing defects. Figure 6 shows a side-by-side comparison of PAUT and TFM/FMC scans of a block with three side-drilled holes; it shows the increased resolution that TFM/FMC affords.

Figure 6. Comparison of (*left*) PAUT and (*right*) total focus method (TFM) / full matrix capture (FMC) scans.



PAUT, TFM/FMC, and TOFD have been used effectively in several industries, including oil and gas (pipeline), aerospace, and nuclear power generation (pressure vessel) industries. Each of these techniques possesses unique and complementary advantages for detecting and sizing discontinuities. Compared with conventional UT, PAUT offers greater sensitivity, coverage, and speed because it provides the ability to rapidly and repeatedly steer, focus, and scan the ultrasonic beam electronically. It also stores the scan in a data file that can be reviewed by others at any time, including offsite after testing is complete. The main disadvantages of PAUT are that the instruments are expensive, operation and data interpretation are difficult, and it can be difficult to achieve good coupling when using larger probes (Ditchburn and Ibrahim 2009). Compared to PAUT, TFM/FMC

offers much greater resolution. However, it is slower than PAUT because of the data processing requirements, and it generates large data files. TFM/FMC is also a newer technique and, as a result, may be less readily available than PAUT. Together, the three UT techniques provide the flexibility needed to conduct testing in the wide variety of joint geometries that may be encountered in existing HSS. More recent instruments combine all three techniques into one machine, which is convenient and cost effective.

1.1.2 Certification of NDT Technicians

It is widely recognized that the effectiveness of NDT relies heavily on the capabilities of the technician. In the United States, the American Society for Nondestructive Testing (ASNT) developed and maintains the most widely used program for the certification and qualification of NDT personnel. ASNT manages both an employer-based certification program and a central certification program. In each program, there are various levels of certification (i.e., Level I to III), depending on the capability and responsibility of the technician. Each level has its own requirements and requires the technician to pass both written and practical examinations for certification.

The employer-based system was established in the 1960s and, because of its broad applicability and flexibility, remains the primary practice for certification in the United States. Governed by Recommended Practice No. SNT-TC-1A (ASNT 2020b), each employer is allowed to develop its own "written practice" under the direction of an experienced Level III NDT technician to qualify and certify in-house NDT technicians. SNT-TC-1A is not a mandatory standard or code, but it includes recommended guidelines for training, experience, and examinations. The objective of SNT-TC-1A was to provide more structure and uniformity to the NDT field, while allowing each employer to tailor its written practice to meet its own specific needs. In 2006, ASNT released ASNT Standard CP-189 as an alternative to SNT-TC-1A. CP-189 is similar to SNT-TC-1A, but it is a mandatory standard with more rigid requirements for training, experience, and testing of NDT personnel. The requirements in CP-189 cannot be altered based on an employer's specific needs, but it is still an employer-based certification program. In contrast to the employer-based certification programs, the ASNT Central Certification Program is administered by an independent third-party certification agency. ASNT certification is based on the US adoption of ISO 9712:2005 and is intended to ensure an

unbiased and consistent certification program that is globally recognized (ASNT 2020b).

AWS D1.5 (AWS 2020a) requires that NDT of nonfracture-critical members be performed by a Level II technician certified in accordance with SNT-TC-1A or its equivalent. For NDT of fracture-critical members, AWS D1.5 requires that the Level II technician work under the direct supervision of an ASNT-certified Level III technician. The Unified Facilities Guide Specification (UFGS) 05 59 20 (DoD 2021), *Fabrication of Hydraulic Steel Structures*, specifies that the Level II technician must be certified in accordance with CP-189, rather than SNT-TC-1A.

1.1.3 Round-Robin Testing and the Reliability of Ultrasonic Testing (UT)

NDT inspections are sensitive to many factors. Two different flaws of the same size can produce different signals, different technicians may obtain different signals on the same flaw even when using the same technique, and the same technician can obtain different signals on the same flaw, especially if the instrument is setup and calibrated a second time (Spencer 2001). Factors that can influence the reliability of an NDT inspection include the size and characteristics of the flaw in question, the material and geometry of the specimen containing the flaw, the NDT technique and procedures, the equipment and its calibration settings, flaw acceptance and decision variables, and operator performance (Singh 2001; Spencer 2001; Gruber and Light 2002; Carboni and Cantini 2012; Kurz et al. 2013).

The technician may be the most significant variable in the process (DoD 2009). Technician performance is usually assumed to depend on skills, knowledge, and experience. However, even skillful technicians who are well trained can miss flaws or report false indications (Stephens 2000; Dymkin and Konshina 2000; Rummel 2004; Fucsok 1998). D'Agostino et al. (2017) provided a comprehensive review of research into the influence of human factors on the reliability of NDT. These authors reported that, in addition to skills, knowledge, and experience, human factors also included the personality, temperament, self-confidence, attitudes, and work practices of the NDT technician. While most studies have focused on the mental and physical condition of the technician, the broader scope of team, environmental, and organizational factors are also important (Enkvist et al. 2000; Carter and McGrath 2013).

The reliability of an NDT technique is established through round-robin testing studies. Round-robin testing studies are designed experiments in which multiple NDT technicians take turns applying an NDT technique and procedure to detect, size, and characterize hidden flaws in manufactured weld specimens. Testing is blind because the technicians have no prior knowledge regarding the location, size, or characteristics of the flaws. The results of round-robin experiments are subsequently analyzed to describe the reliability and effectiveness of the NDT technique, depending on the purpose and motivations of the study. Motivations for undertaking round-robin experiments may include demonstrating the reliability of a procedure or technique, comparing the effectiveness of two or more NDT techniques or procedures, or determining inspection intervals as part of a damage tolerant design program.

Round-robin experiments have been used to investigate the reliability and effectiveness of UT since at least 1965, when the US Pressure Vessel Research Committee (PVRC) studied the ability of a UT procedure to detect, locate, and size flaws in nuclear reactor steel (Crutzen 1985). Three of the steel plates from the PVRC study were later transferred to the Commission of the European Communities Joint Research Center, where the Plate Inspection Steering Committee (PISC) was formed in 1976 to conduct a round-robin study. That study, which is now known as PISC-I, investigated the same set of procedures as the PVRC (OECD 1986). Crutzen (1985) reported that neither the PVRC study nor the PISC-I study produced any useful results, but Carvalho et al. (2006) reported that the PISC-I study demonstrated the need to improve NDT technology. In 1980, a growing emphasis on in-service inspection of nuclear reactors made it increasingly important to understand the reliability of NDT techniques and procedures (OECD 1986). The Programme for the Inspection of Steel Components carried out two studies, known as PISC-II (1981–1984) and PISC-III (1985–1988), which involved at least 34 teams of technicians from 10 European countries (Crutzen 1985; OECD 1986).

In the 1970s, the US Air Force (USAF) and the National Aeronautics and Space Administration (NASA) began round-robin testing of NDT techniques to support a transition to damage-tolerant designs (Singh 2001). Today, aircraft components are designed for regular inspection using NDT, and round-robin experiments are used to assess NDT capabilities with respect to those individual components (Singh 2001; Grandt 2011). Within the DoD, MIL-HDBK-1823A (DoD 2009) provides guidance on

conducting probability of detection (POD) studies to assess the capabilities of an NDT system and establish NDT inspection procedures for components of flight propulsion systems, airframes, and ground vehicles. According to DoD (2009), the goal of these studies is to estimate POD as a function of a flaw characteristic, a, which typically represents flaw length. The $a_{90/95}$ is the lower bound of the 95% confidence interval on the target size a with a POD = 0.90 and represents the largest crack that might be missed during an inspection. The results of POD studies provide a basis for determining the inspection interval for aircraft components (DoD 2009; Cherry and Knott 2022; Knott and Schubert Kabban 2022b).

There are two primary types of POD studies (DoD 2009; Cherry and Knott 2022). These include $\hat{a}$ versus a studies and hit/miss studies. The major differences between the two approaches are the type of data collected and the statistical methods used to analyze those data. The $\hat{a}$ versus a studies collect continuous data on the strength of the UT signal being reflected back to the probe and analyze those data using linear regression (DoD 2009; Cherry and Knott 2022). Hit/miss studies collect binary data on whether or not a flaw was detected and analyze those data using logistic regression (Agresti 2013; Knott and Schubert Kabban 2022b). Hit/miss methods may be used to analyze $\hat{a}$ versus a data if the observations are first converted to binary by adopting a critical threshold signal strength to identify flaws. However, the hit/miss and $\hat{a}$ versus a methods will produce different results when applied to the same dataset. The conclusions from POD studies are also sensitive to the methods used to construct confidence intervals (Knott and Schubert Kabban 2022a). Finally, it is widely recognized that POD study results may be very uncertain as a result of human factors and the failure or inability to control other sources of variability during round-robin experiments. This may limit the utility of POD study results (Keprate and Ratnayake 2015; Knott and Schubert Kabban 2022b).

When evaluating the capability of an NDT technique, it is important to mimic the actual inspection process as closely as possible (Carvalho et al. 2006; DoD 2009). This means controlling those factors that will be controlled during the inspection and allowing those factors that will not be controlled to vary (DoD 2009; Cherry and Knott 2022). The purpose and motivation for the round-robin experiment dictate which factors are controlled. For example, in the context of an NDT reliability study for a damage-tolerant design, it is desirable to define the POD in terms of a single dimension and to minimize uncertainty in the POD to maximize the

inspection interval. This can be achieved by controlling the inspection technique and procedures, material type, weld configuration, specimen geometry, and plate thicknesses. In other cases, the purpose and motivation of the study may justify representing the variability in these factors. For example, the present study is focused on understanding the applicability of PAUT to HSS in general, and the weld specimens represent a variety of typical HSS joint geometries. Emphasis is placed on characterizing the uncertainty in estimates of flaw length and height because this information is needed to derive partial safety factors for FFS analysis.

Numerous authors have described the practical challenges of round-robin studies. They are expensive, time consuming, and logistically difficult (Singh 2001; DoD 2009; Keprate and Ratnayake 2015; Virkkunen et al. 2022; Knott and Schubert Kabban 2022b). Specimens must be designed and manufactured to contain a sufficient number of flaws that range in size from almost always missed to almost always detected. Technicians who are willing and able to participate in the round-robin experiment must be recruited and brought to a central location. The participation of technicians and researchers in one or more round-robin events needs to be coordinated and scheduled. In addition to practical issues, there are a number of other well-known factors that may hinder the interpretation of study results. For example, it is very difficult to mimic real testing conditions in round-robin studies. Round-robin tests are typically conducted in closed, climate-controlled environments with good lighting that do not represent environmental conditions in the field, which can sometimes be harsh. The weld specimens are often smaller mock-ups designed to represent larger structures. As such, they are easier to handle, and technicians may be able to access the welds more easily than the welds of in situ structures. In addition, the NDT technicians know that their work will be evaluated during a round-robin experiment, and as a result, they may be more highly motivated to perform well during the testing than in the field (Carvalho et al. 2006). These factors suggest that round-robin studies may tend to overstate the capability of NDT techniques in practice.

There are many examples of round-robin studies in the literature. Several examples are listed in Table 1 so the reader can compare the scope and size of the present study with the scope, size, and findings of other round-robin studies. The present study is listed as ERDC 2024 in the last row of Table 1. Quantitative assessments of NDT from any one round-robin experiment depend on the context of the study and are difficult to extrapolate

(Carvalho et al. 2006), so those are not reported here. However, the studies listed in the table either informed the approach taken in the present study or reported findings similar to those of the present study.

Spencer (1996a, 1996b) studied visual inspection of aircraft welds using round-robin experiments to estimate POD as a function of crack length. The author found substantial variations in POD curves among technicians that could not be explained by recent inspection experience. He also found that, for a population of cracks taken from many areas of the aircraft and from many different types of structures, crack length could not be used to explain the variation found in detection rates.

Table 1. Examples of round-robin testing experiments in the literature.

Citation	Context/Material	Technique	Technicians	Specimens	Flaws
Spencer 1996a, 1996b	Visual inspection of aircraft welds	Visual	12	10	64
Fucsok 1998	Evaluate influence of human factors on manual ultrasonic inspection of manufactured specimens	PE	24	1	5
Gruber and Light 2002	Structural steel; beam to column weld geometries	PE	1	12	17
Shaw 2002	Structural steel connections	PE	15	12	13
Carvalho et al. 2006	Compare manual and automatic inspection in girth weld beads of API X70 steel pipeline	PE, TOFD	5	24	90
Schneider and Bird 2009	Flat carbon steel plates with butt welds	PAUT, PE	10	8	40
Kurz et al. 2013	Nuclear facility steels (austenitic, ferritic, dissimilar, cladded)	PAUT	3 teams	36	>128
EPRI 2018; Jacob et al. 2018	Cast austenitic stainless steel (CASS) used in nuclear plants	PE	7	20	—
Connor et al. 2019	Complete joint penetration bridge welds	PAUT, TOFD, PE	11	—	19
Boone et al. 2019	Automatic ultrasonic inspection of bridge welds	PAUT	4	6	14
Choi et al. 2022	Tube weld steels (American Society of Mechanical Engineers [ASME] SA210Gr.A1, ASME TP304H)	PAUT	6 teams of 2	30	109
ERDC 2024	Specimens representing HSS joint geometries	PAUT	18	20	68

Gruber and Light (2002) investigated the capability of AWS D1.1 code procedures when applied to beam-to-column weld geometries and the potential for supplemental procedures to increase the POD and facilitate flaw location, classification, and sizing. These authors found that when using the amplitude (i.e., decibel) drop method, large flaws were just as likely to be missed as small flaws and that weld crowns, backing bars, and web access holes limited the ability to size flaws. In contrast to Spencer (1996a), these authors found a positive relationship between POD and the knowledge, skills, and experience of NDT technicians.

Shaw (2002) investigated the capability of UT to detect and size flaws in structural steel connections using the AWS D1.1 structural welding code. Indication ratings (i.e., decibel values) varied from one technician to another by as much as 15 dB for a given flaw. On average, technicians missed 25% of known flaws, and large flaws were just as likely to be missed as small flaws. They also found that the false positive (FP) rate of 16% was reduced to 7% when specimens with backing bars were excluded. With respect to flaw sizing, technicians located flaw start and stop locations within 0.24 in. (6 mm). Flaw length was estimated within ±0.24 in. (6 mm) 65% of the time and within 0.51 in. (13 mm) 85% of the time.

Carvalho et al. (2006) conducted round-robin experiments on 24 girth welds of a steel pipeline containing flaws characterized as lack of fusion (LOF) and lack of penetration (LOP). Results were used to estimate POD curves for manual (i.e., PE) and automatic (i.e., PE, TOFD) ultrasonic techniques as a function of flaw length. Results showed that automatic techniques were superior to manual techniques. This was attributed to the reduced influence of human factors in automatic techniques. The effect of human factors was further demonstrated by variability in the POD curves and flaw size estimates across technicians. Technicians tended to overestimate the length of all flaws, which ranged in length from 0.12 to 0.79 in. (3 to 20 mm). Proportional errors in flaw length were much greater for shorter flaws than for longer flaws. The authors noted that automatic inspection is still subject to human factors because humans are required to set up the equipment, calibrate the instrument, and interpret the results.

Schneider and Bird (2009) compared the ability to detect and size flaws using PAUT and PE (i.e., manual UT) by conducting round-robin tests on eight steel test blocks containing 40 planar and volumetric flaws. Fifteen technicians participated in the study, although data from three of the

technicians were discarded because they performed so poorly. The authors used a balanced experimental design with respect to flaw type, size, and location within each test block. Analysis of variance (ANOVA) was used to identify factors influencing the ability to size flaws while accounting for differences in procedure and equipment, operator, operator experience and qualifications, wall thickness (including asymmetry in the block), weld shape, and flaw size, type, and location. Results showed that flaw height, operator, and, to a lesser extent, plate thickness were significant factors. These authors also found that error in flaw height increased with flaw height and that there was a tendency to overestimate the height of shorter flaws and to underestimate the height of taller flaws. Errors in estimates of flaw height by technicians with certain phased array defect sizing credentials were about 20% less than those without the credentials. Errors in estimates of flaw height by technicians with the height sizing credentials plus more than five years of experience were about 40% less than those without the credentials.

Kurz et al. (2013) conducted round-robin experiments using an $\hat{a}$ versus a study approach to estimate POD and quantify uncertainty in flaw size estimates for different types of steels used in nuclear power plants (i.e., austenitic, ferritic, dissimilar, and cladded). Although NDT is an essential part of the construction and maintenance of nuclear power plants in Germany, there is a lack of information regarding the influence of different flaw types, different steels, and testing conditions on POD and flaw size estimates. EPRI (2018) and Jacob et al. (2018) also reported on a round-robin experiment sponsored by the Electric Power Research Institute (EPRI) to quantify the capability of NDT techniques applied to cast austenitic stainless steel (CASS) used in nuclear power plants. These authors found that no NDT technique or technician was able to demonstrate a detection rate of 80% with less than a 20% false call rate. They also noted a pronounced tendency to undersize flaw length and height, particularly with respect to larger flaws.

Connor et al. (2019) and Boone et al. (2019) reported on round-robin experiments to quantify the reliability of PAUT, TOFD, and PE applied to bridge welds. The authors found that the critical flaw size of bridge welds could be developed using FFS analysis, but a large amount of variability was possible when weld inspections were performed using the AWS D1.5 (AWS 2015) PE and PAUT scanning procedures. These authors found that the acoustic properties of bridge steels vary widely and may not be

isotropic. HSS are constructed of steels with specifications similar to those used in bridges. Finally, these authors found that there was a need to independently administer practical exams in addition to any employer-based certifications.

Choi et al. (2022) conducted round-robin experiments using flawed tube weld specimens representing tubes in the boilers of nuclear power plant facilities. The objective of this study was to assess the capabilities of PAUT as a substitute for RT. POD curves were estimated for different materials (e.g., ferritic steel, austenitic stainless steel, and dissimilar metals) and flaw types (i.e., volumetric, planar). The authors showed that PAUT was superior to RT. Overall, there was a pronounced tendency to overestimate flaw length.

1.2 Objectives

The research gap addressed in this study is the lack of information on the capabilities of PAUT with respect to detecting, sizing, and characterizing flaws in existing HSS. Two recent studies investigated the capabilities of PAUT with respect to bridge welds (Connor et al. 2019; Boone et al. 2019). Bridges are similar to HSS in terms of both the materials used in construction and the applicable set of AWS codes that govern NDT. However, HSS have more complex joint geometries, including thicker plates, thickness transitions, and skewed joints, that may make NDT more difficult. The round-robin experiments described in this paper were carried out to quantify uncertainty in the ability to detect, size, and characterize flaws. The data were analyzed to quantify POD, uncertainty in estimates of flaw length and height, and uncertainty in flaw characterization. Safety factors were derived from characterizations of uncertainty in flaw size estimates. This report demonstrates the application of safety factors in examples of FFS analysis of HSS. This study concludes with specific recommendations based on study results.

1.3 Approach

The US Army Engineer Research and Development Center (ERDC) designed 12 steel specimens, referred to here as the *ERDC specimens*, to represent HSS weld and joint configurations (Appendix A). These included a variety of butt joints, corner joints, and T-joints, including skewed joints and thickness transitions. An additional eight specimens, referred to here as the *AWS specimens*, were procured as part of an off-the-shelf kit

intended for UT training specific to the bridge industry and the bridge welding code (AWS D1.5, AWS 2020a; Appendix B). The 20 specimens contained 21 welds and 68 flaws, representing a variety of types, sizes, and orientations distributed at known locations. Flaws were embedded in the welds of test specimens. Flaw categories included planar, volumetric, and laminar flaws. The planar flaw category included three subcategories: LOF, LOP, and cracks (CRKs). The CRK subcategory included root cracks (ROCRKs), base metal cracks (BMCRKs), transverse cracks (TRCRKs), centerline cracks (CLCRKs), and toe cracks (TOCRKs). The volumetric flaw category included two subcategories: porosity (POR) and slag (SLAG). The laminar flaw category included one subcategory: laminations (LAM). The locations of flaws and their dimensions were certified by the manufacturer and verified independently through blind testing. Verification testing was performed by an ASNT Level III certified NDT technician who confirmed the location and dimensions of each flaw. This technician also confirmed that all flaws were Class A rejectable per AWS D1.1 (AWS 2020c) under static load and per AWS D1.5 (AWS 2020a) for primary members subject to tensile stress.

The ERDC research team worked with experts in industry and academia to develop, test, and validate UT inspection procedures for detecting, sizing, and characterizing flaws in the steel specimens. Eighteen NDT technicians certified in accordance with Recommended Practice No. SNT-TC-1A (ASNT 2020b) were recruited to participate in a series of round-robin experiments to detect, size, and characterize the flaws. These technicians were primarily from the oil and gas industry, rather than from the structural industry, because PAUT and flaw height sizing are already well-established techniques in that industry. PAUT was only recently added to the AWS codes that govern testing in the structural industry, and AWS codes do not require height sizing. Technicians were selected from among a pool of 27 candidates who had previously submitted to a prequalification test that consisted of written and practical components. Prior experience working with HSS was not considered when recruiting technicians, and with one or two exceptions, none of the candidates described prior experience working on HSS or for USACE. All technicians who participated in prequalification or round-robin testing were paid by their employers. Each of the employers was subcontracted to ERDC through a contract with Michael Baker International.

Round-robin testing took place over three one-week periods in the spring of 2022, with up to eight technicians participating in any given week. Round-robin testing was conducted in a large conference room at the Coastal and Hydraulics Laboratory (CHL) at ERDC in Vicksburg, Mississippi. Each technician applied a consistent set of NDT procedures that were developed and validated specifically for this research project. These procedures included manual rastering (i.e., rotation of the probe with respect to the weld axis), line scans from multiple index offsets, scanning from all faces, and the use of smaller probes. Technicians operated at their own pace, with most technicians scanning the full set of 21 welds within a five-day period. Each time a discontinuity was detected, the technician recorded the location of the beginning and end of each flaw in each dimension, estimated the length and height of each flaw using the decibel drop method or diffraction techniques, and characterized its type. The research team reviewed each data form when it was submitted and resolved any questions or ambiguous data entries before allowing the technician to proceed to the next specimen.

Round-robin test data were analyzed to (1) model the probability of flaw detection, (2) model the probability that reported indications corresponded to actual flaws, (3) calculate confidence bounds on estimates of flaw length and height, and (4) assess the accuracy with which technicians can characterize flaws. Logistic regression models were fit to round-robin testing data to model the POD and assess the influence of various flaw and specimen characteristics on flaw detectability. These analyses demonstrated the applicability of UT techniques to HSS in general and helped identify the specific conditions under which UT may be more or less effective on HSS. Logistic regressions were also fit to model the probability that reported indications corresponded to actual flaws. Information on the reliability of reported flaw indications may help HSS managers distinguish between true positive (TP) and FP indications in the field and to recognize signs that an indication may require independent verification. Errors in flaw size estimates were analyzed to derive partial safety factors for use in Engineering Critical Assessments. This will enable HSS managers to account for uncertainties in flaw size estimates when conducting FFS analysis. The accuracy with which NDT technicians can characterize flaws using UT techniques was also assessed. Information on the character of flaws is also needed for FFS analysis.

This report is organized as follows. Chapter 2 describes the specimens used in round-robin testing and summarizes information on the flaws. Each specimen represents one of eight joints that are described in Appendix A. As-built drawings and photographs of the specimens can be found in Appendix B and Appendix C. Chapter 3 describes the development and validation of procedures for three NDT techniques (i.e., PAUT, TFM/FMC, and TOFD). These procedures were initially discussed at an expert workshop (the minutes of which are provided in Appendix C) and subsequently developed by ASNT Level III technicians. Final procedures are provided in Appendix E. Chapter 4 describes the prequalification of NDT technicians, the round-robin experiments, and study results. Chapter 5 provides examples of how these results can be incorporated into FFS analysis. Appendix F presents detailed calculations for the FFS analysis examples. Chapter 6 describes the general conclusions and practical implications of this study with respect to inspection and maintenance of HSS.

2 The Design and Manufacture of Weld Specimens

The ERDC research team initially designed a set of 12 welded steel specimens to represent joint geometries and weld configurations encountered in HSS that were constructed in the 1960s and 1970s. An additional eight specimens were subsequently procured from an off-the-shelf test kit that was designed specifically for the bridge industry. These two groups of specimens are described in Table 2 and are subsequently referred to as the ERDC specimens and the AWS specimens, respectively.

2.1 The Design of US Army Engineer Research and Development Center (ERDC) Weld Specimens

In the course of designing the ERDC specimens, the research team reviewed construction plans from a number of in-service steel structures constructed in the 1960s and 1970s to identify common joint geometries. These structures included miter gates, sector gates, vertical lift gates, tainter gates, dewatering bulkheads, and needle girders. Eight representative HSS joint geometries were identified. Appendix A provides descriptions of these joints and the drawings on which they were based. This information was distributed to select members of the USACE Structural Community of Practice (COP), who provided additional insight into which joints were prone to issues during service and inspection, weld procedures, and joint detailing. One of the reviewers also recommended that the test specimens include a butt joint with a square groove weld similar to the welded detail in the maintenance stop log that failed at Coffeeville Lock and Dam in 1994 (USACE 2009).

Other characteristics of the 12 ERDC specimens were selected after determining the joint type (e.g., butt joint, T-joint, or corner joint) for each specimen. These included joint configuration (e.g., plate thickness, connection angle, joint prep, number of flaws), flaw category, and flaw size. The range and distribution of each variable were determined by the research team in consultation with the specimen fabricator and experts from the Structural COP and academia. An approximate form of Latin hypercube sampling was used to obtain a distributed and nearly random set of values for each variable by partitioning a lognormal distribution function into a series of uniformly sized bins. The exact value of each variable was then determined based on the judgement and experience of the research

team. Final adjustments to flaw size (i.e., length and height) and plate thickness were made at the recommendation of the manufacturer. The minimum flaw length and height were 0.06 in. and 0.04 in., respectively. Material thickness was rounded to the nearest 1/8 in.

Table 2 provides the final specimen matrix. The specimen identifier is a unique code assigned by ERDC. Specimen ERDC 007 contains two welds, so this specimen is represented twice: as ERDC 007A and as ERDC 007B. Joint type describes the basic shape of the joint (e.g., butt, corner, T). The column labeled AWS Joint identifies the AWS joint designation that is represented by each of the ERDC specimens. Joint designations are described in AWS D1.5 (AWS 2020a), Clause 4, and in Figure 27. The column labeled HSS Joint lists which one of the eight representative HSS joint geometries (labeled A–H and described in Table A-1) is represented by each of the ERDC specimens. The AWS specimens are not associated with AWS joint designations or representative HSS joint geometries.

All specimens were fabricated by FlawTech, a company based in Concord, North Carolina, that specializes in manufacturing flawed specimens for NDT and training purposes. The specimens were fabricated from ASTM A36 steel and welded in accordance with AWS D1.1 (AWS 2020c). As is typical in older HSS, no weld access holes were included, and welds were not ground flush. The maximum ratio of flaw length to weld length in any given specimen was less than 0.5. The width of the specimens (perpendicular to the weld axis) was selected to allow inspection in at least the first and second legs of the sound path. During testing, the technicians were given access to all four sides of the specimens. An additional eight specimens were procured as part of an off-the-shelf kit intended for UT training specific to the bridge industry and the bridge welding code (AWS D1.5; AWS 2020a). After fabrication, all of the specimens were coated with 8–12 mil of a three-coat vinyl paint system (System No. 5-E-Z). This coating was to prevent the specimens from rusting, and its effect on UT was judged to be similar to that of vintage coatings. However, this coating is more representative of those used on newer HSS than on vintage HSS. It has better adherence than vintage coatings and, therefore, does not replicate the presence of chipping and cracking that might interfere with UT of vintage HSS. A handle was installed on all specimens weighing more than 50 lb to allow for easier handling.

Table 2. Specimen matrix for custom-built and off-the-shelf test specimens.

Specimen Identifier	Joint Type	AWS Joint	HSS Joint	Weld Length (in.)	Weight (lb)	Plate 1 Thickness (in.)	Plate 2 Thickness (in.)	Joint Angle (deg)
ERDC 001	Butt	B-U2	A	12	30	0.625	0.625	180
ERDC 002	Butt	B-U3b	B	6	350	4	4	180
ERDC 003	Butt	B-L1b	B	12	25	0.625	0.50	180
ERDC 004	Butt	B-U2	C	12	30	0.64	0.625	160
ERDC 005	Butt	B-U3b	D	4	30	1	1	90
ERDC 006	Butt	B-U2	D	7	15	0.5	0.5	35
ERDC 007A	Butt	B-U3b	E	7.625	70	WT9 × 65 (*tf* = 1.2)	WT9 35.5 (*tf* = 0.81)	180
ERDC 007B	Butt	B-U2	E	8.4375	70	WT9 × 65 (*tw* = 0.67)	WT9 × 35.5 (*tw* = 0.495)	180
ERDC 008	T	TC-U5b	F	6	125	1.5	1.5	90
ERDC 009	T	TC-U4b	G	10	55	0.75	0.75	90
ERDC 010	T	TC-U4b	G	6	15	0.5	0.5	35
ERDC 011	Corner	TC-U5b	H	6	120	1.875	1.875	90
ERDC 012	Corner	TC-U4b	H	6	70	1.375	1.375	110
AWS-BK-01	Butt	—	—	12	30	0.75	0.75	180
AWS-BK-02	Butt	—	—	12	50	1.0	1.0	180
AWS-BK-03	Butt	—	—	12	35	1.0	0.5	180
AWS-BK-04	Butt	—	—	12	30	0.75	0.75	180
AWS-BK-05	T	—	—	12	30	1.0	0.75	90
AWS-BK-06	T	—	—	6	25	0.75	0.5	45
AWS-BK-07	Butt	—	—	6	15	1.0	0.5	180
AWS-BK-08	Corner	—	—	6	15	0.75	0.75	90

Note: WT refers to a shape cut from a wide flange beam, *tf* refers to flange thickness, and *tw* refers to web thickness.

Figure 7 shows representative specimens. The faces and coordinate axes of each specimen were labeled to establish consistency in testing and reporting of results. The specimen coordinate system was established to match AWS conventions, with the *x*-axis perpendicular to the weld axis, the *y*-axis parallel to the weld axis, and the *z*-axis positive from face A to face B. Figure 8 illustrates the conventions used in labeling the faces and coordinates of each joint type. Appendix B contains as-built drawings and photos of all 12 ERDC specimens. Appendix C provides as-built drawings of AWS specimens. Table 2 contains information on dimensions and joint types.

Figure 7. Representative welded steel specimens with faces and coordinates labeled.

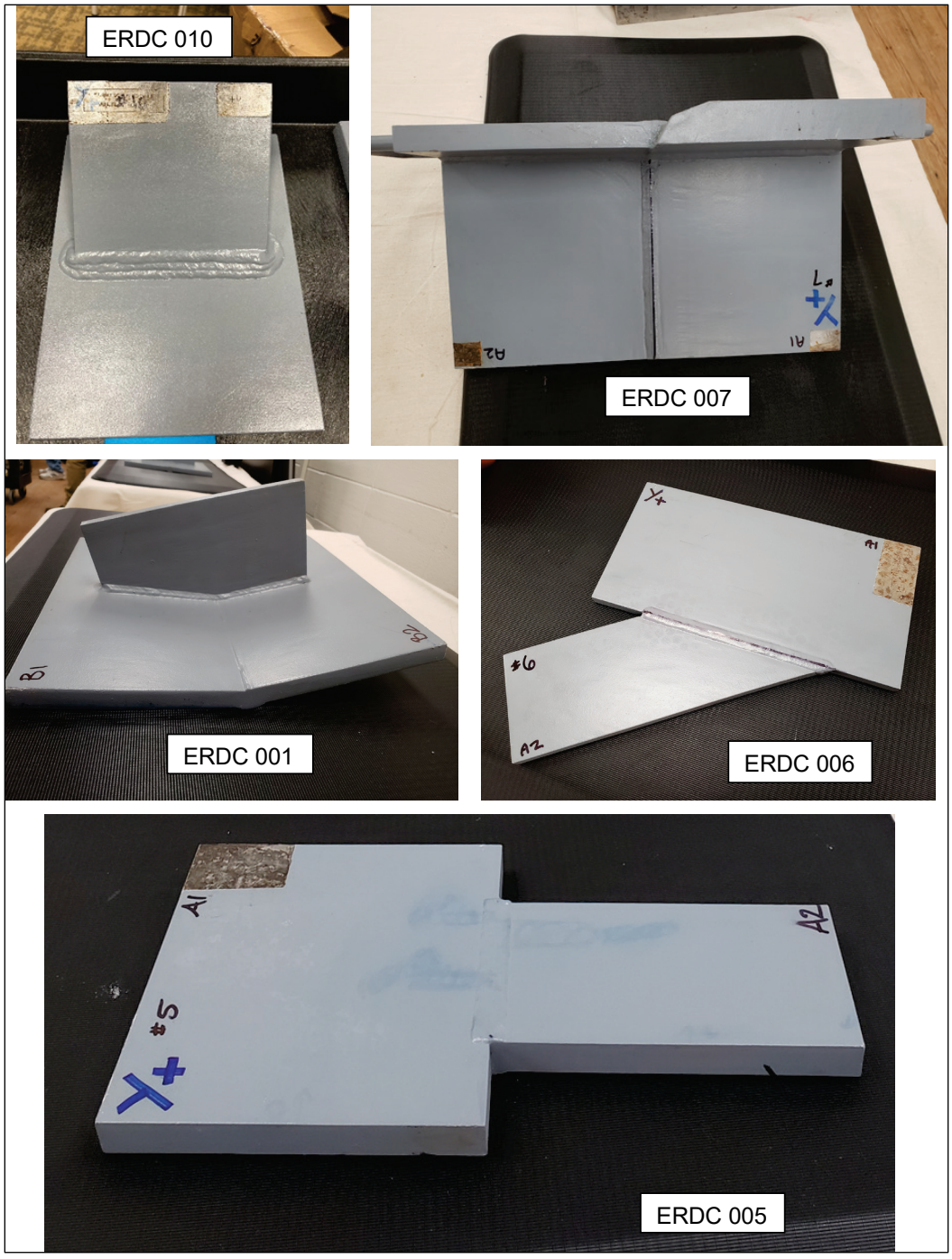
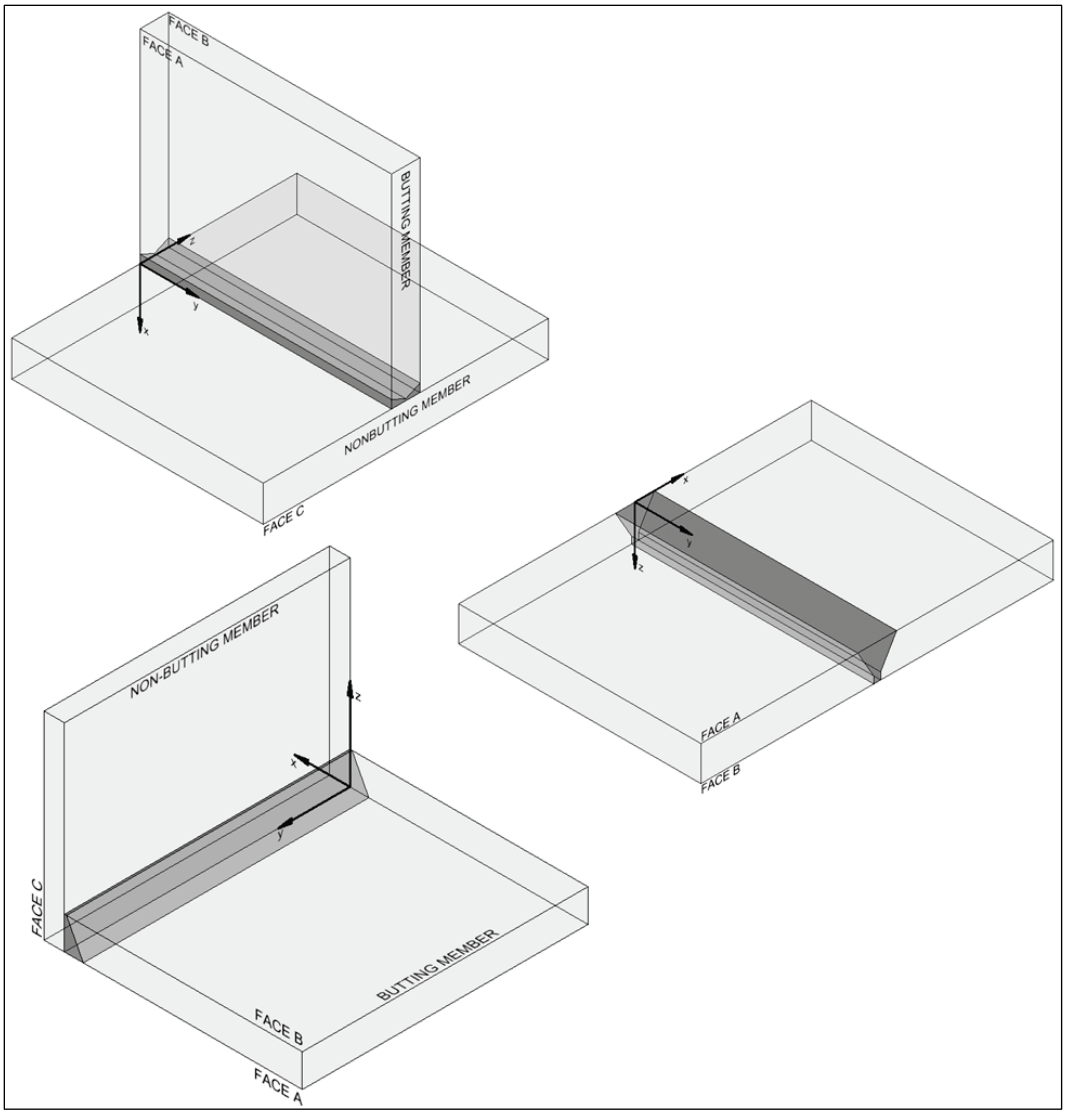


Figure 8. Conventions used in labeling faces and axes of three different joint types.



In total, 68 flaws were distributed within the 20 specimens. Table 3 contains descriptions of each flaw. Several different flaw categories were included in the specimens to provide an opportunity to assess the ability of NDT technicians to characterize flaws and to detect and size different flaw categories. Each flaw can be classified by category and subcategory. Flaw categories include planar, volumetric, and laminar. Flaw subcategories describe subgroups within each flaw category and are defined in Table 3 per AWS (2019). Planar flaw subcategories include LOF, LOP, and CRK. Volumetric flaw subcategories include SLAG and POR. Laminar flaws include one subcategory: LAM. Laminations are a weakness generally aligned parallel to the worked surface of a metal and occur during manufacture; they should not be confused with lamellar tears, which are step-like cracks that

develop either during welding or while a component is in service (AWS 2020b). Figure 9 shows the distribution of flaws across flaw subcategories. All flaws would be considered Class A rejectable per AWS D1.1 (AWS 2020c) for cyclically loaded nontubular members in tension. The character, size, and location of flaws are documented in as-built drawings and UT test reports from the manufacturer. However, no information about the location of flaws within each specimen is provided in this report to preserve the ability to use those specimens in future blind testing. The research team verified the as-built drawings using manual contact PE inspection. No RT or destructive examination of the specimens was performed.

Figure 9. Distribution of flaws by flaw subcategory.

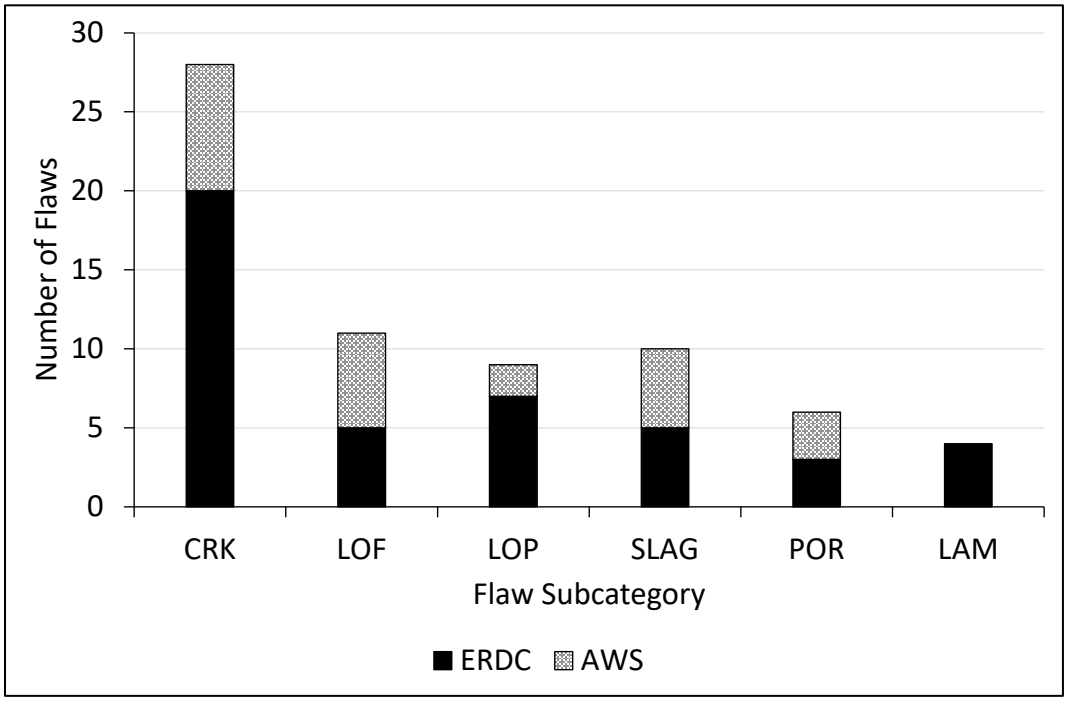


Table 3. Flaw categories, subcategories, definitions (AWS 2020b), and dimensions in US Army Engineer Research and Development Center (ERDC) and American Welding Society (AWS) specimens.

Flaw Category		Flaw Subcategory and Definitions from AWS (2019)	Length (in.)	Height (in.)	Area (in.²)	Orientation (deg)
Planar	Crack: a fracture-type discontinuity with a sharp tip and high ratio of length and width to opening displacement.	Base metal crack: a crack located in the base metal.	0.3	0.13	—	8
			1.37	0.16	—	−5
			1.47	0.31	—	27
		Centerline crack: a crack located in the center of the weld.	0.46	0.11	—	0
			1.72	0.25	—	0
			0.24	0.09	—	21
			2.59	0.38	—	−12
			1.27	0.34	—	12
			0.87	0.19	—	−28
			1.02	0.39	—	−10
		Centerline / base metal crack	0.85	0.45	—	0
		Root crack: a crack located in the weld root.	0.08	0.07	—	0
			0.40	0.12	—	0
			1.27	0.15	—	0
			0.41	0.10	—	0
			0.55	0.23	—	0
		Toe crack: a located in the weld toe.	0.59	0.09	—	0
			0.46	0.25	—	0
			3.35	0.51	—	14
			0.54	0.141	—	0
			0.61	0.12	—	0
			0.76	0.36	—	1
			0.19	0.16	—	0
			1.96	0.14	—	−18
		Transverse crack: a crack with its major axis oriented approximately perpendicular to the weld axis.	0.16	0.16	—	0
			0.91	0.16	—	0
			1.05	0.14	—	0
			0.37	0.18	—	0
	Lack of penetration: a joint root condition in a groove weld in which weld metal does not extend through the joint thickness.		1.31	0.16	—	0
			0.67	0.093	—	0
			0.58	0.19	—	0
			3.00	0.16	—	0
			0.65	0.044	—	0
			0.29	0.10	—	0
			2.79	0.16	—	0
			1.02	0.54	—	0
			1.47	0.16	—	0
	Lack of fusion: a weld discontinuity in which fusion did not occur between the weld metal and the fusion faces or the adjoining weld beads.		2.11	0.20	—	0
			0.17	0.04	—	0
			0.82	0.16	—	0
			0.57	0.07	—	0
			0.27	0.00	—	0
			0.86	0.23	—	0
			0.39	0.15	—	0
			0.46	0.19	—	0
			4.52	0.21	—	0
			1.47	0.18	—	0
			0.64	0.15	—	0
Volumetric		Porosity: cavity-type discontinuities formed by gas entrapment during solidification or in a thermal spray deposit.	0.49	0.14	—	0
			0.68	0.17	—	0
			0.98	0.21	—	0
			2.77	0.14	—	0
			1.37	0.26	—	0
		Porosity (subsurface): see porosity.	0.28	0.20	—	0
		Slag inclusion: a nonmetallic byproduct of the mutual dissolution of flux with nonmetallic impurities in welding and brazing processes.	1.29	0.12	—	0
			0.56	0.21	—	0
			1.09	0.13	—	0
			0.77	0.08	—	0
			2.37	0.05	—	0
			6.98	0.11	—	0
			0.28	0.08	—	0
			0.62	0.21	—	0
			1.16	0.097	—	0
			0.60	0.08	—	0
Laminar		Lamination: a type of discontinuity with separation or weakness generally aligned parallel to the worked surface of a metal.	0.71	0.00	0.5	0
			1.41	0.00	2	0
			1.01	0.00	1.05	0
			0.35	0.00	0.12	0

Figures 10 and 11 summarize the distribution of flaw lengths and heights. Flaw lengths within the test specimens ranged from 0.08 in. to 6.98 in., with the majority of the flaws between 0.5 in. and 1.5 in. in length. The average flaw length was 1.11 in. Flaw heights ranged from 0.00 in. (for LAM) to 0.54 in., with the majority of the flaws between 0.15 in. and 0.25 in. in height. The average flaw height was 0.18 in. The aspect ratio, defined as the ratio of flaw height to flaw length, ranged from 0.00 and 0.98.

Figure 10. Distribution of flaws by length.

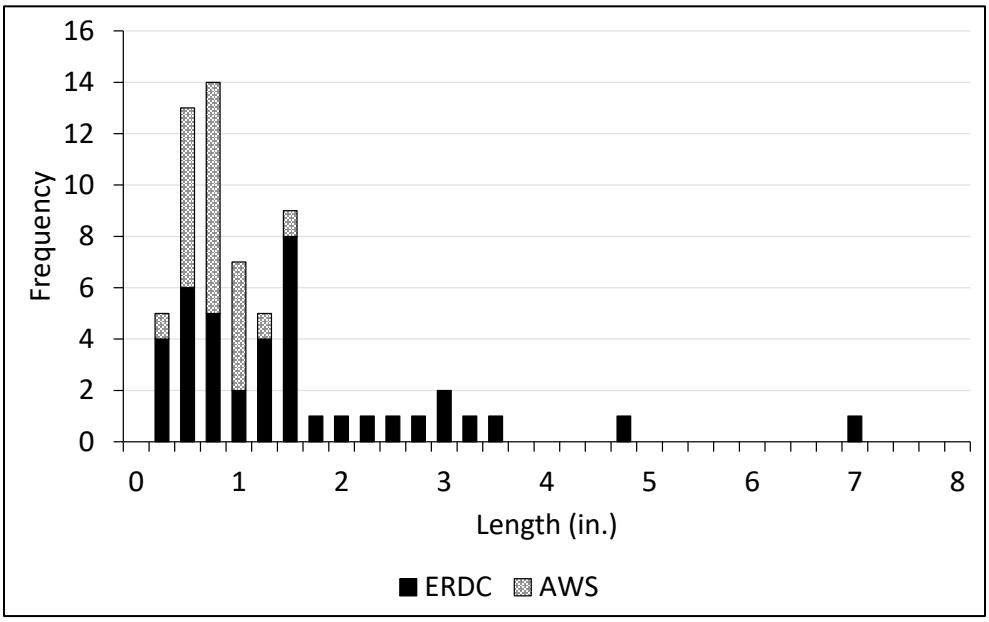
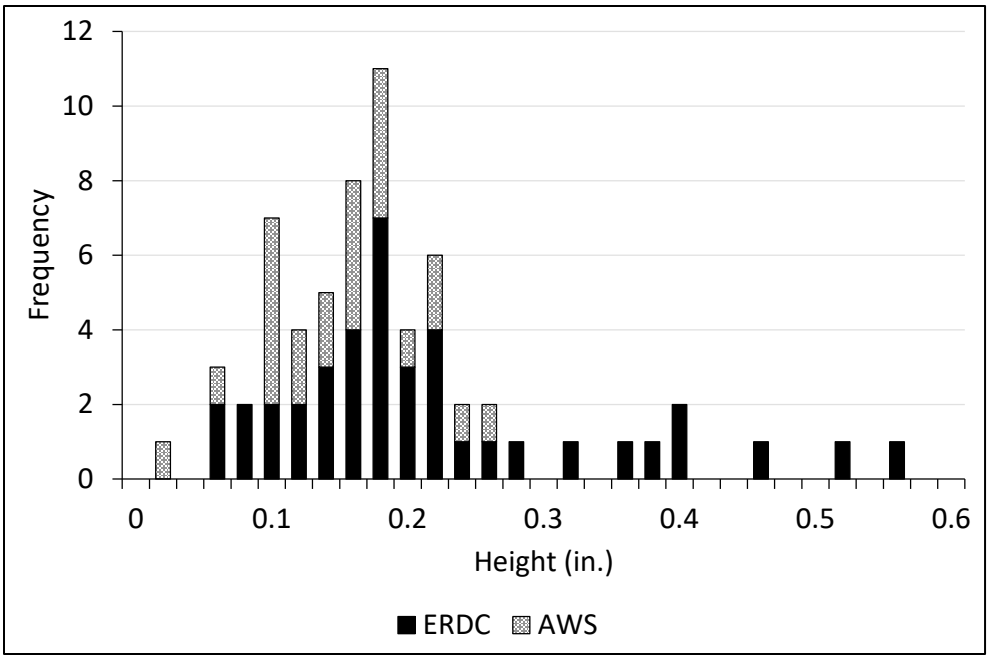


Figure 11. Distribution of flaws by height.

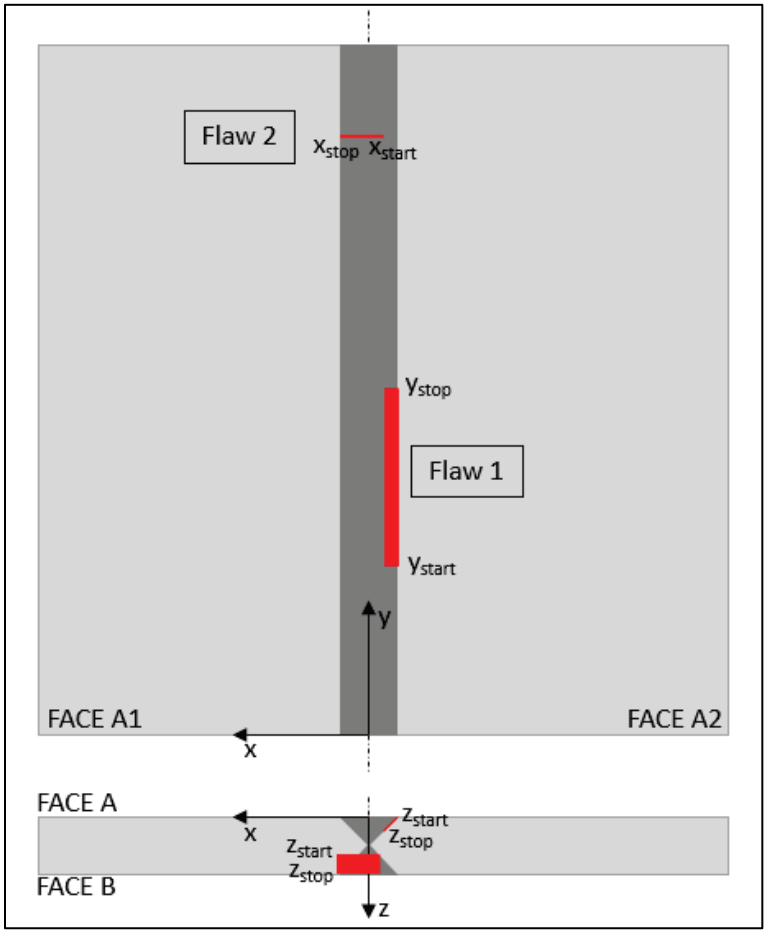


To compare reported indications with true flaws during testing, an answer key was developed for each specimen. This answer key included the start and stop locations for each flaw in all three directions, the flaw category, and the flaw dimensions (i.e., length and height). The specimen coordinate system was established to match AWS convention with the x-axis perpendicular to the weld axis, the y-axis parallel to the weld axis, and the z-axis positive from face A to face B (Figure 12). For flaws oriented parallel to the weld axis, the flaw length was calculated as $l_y = y_{stop} - y_{start}$. For flaws oriented perpendicular to the weld axis (e.g., TRCRKs), flaw length was calculated as $l_x = x_{stop} - x_{start}$. For flaws oriented at an angle between 0° and 90° to the weld axis (i.e., skewed), the length was calculated as

$$l_{xy} = \sqrt{(x_{stop} - x_{start})^2 + (y_{stop} - x_{start})^2}.$$

For all flaws, height was calculated as $h = z_{stop} - z_{start}$ regardless of its orientation relative to the z-axis (i.e., tilt).

Figure 12. Illustration of flaw coordinates.



2.2 Verification of the Location and Size of Flaws in Weld Specimens

The location and size of each flaw in ERDC's weld specimens was independently verified by a member of the research team who is an ASNT Level II and III certified NDT technician. This team member was not involved in specimen development. Therefore, all verification testing was performed blind, without the technician having any prior knowledge of the locations and sizes of flaws in the specimens. Manual contact PE UT was used in verification testing. Calibration of the manual PE instrument and verification testing were performed in accordance with AWS D1.1 (AWS 2020c) and AWS D1.5 (AWS 2020a). Detection and acceptance or rejection were accomplished using a standard AWS transducer (2.25 MHz) with a refracted angle of 45°, 60°, or 70°.

A transfer correction factor of 10 dB was added to sensitivity to compensate for the 5-E-Z vinyl coating. This factor was based on a sensitivity calibration performed on the 0.06 side-drilled hole in an International Institute of Welding (IIW) block prior to coating and again after coating. Sizing was accomplished using a 5 or 7.5 MHz shear wave transducer. However, a 2 MHz creeping wave transducer was used for toe cracks. The 6 dB drop method was used to identify the limits of the flaws for length measurements. Time-based techniques were used for height sizing. A high frequency transducer (7.5 MHz) with a refracted angle of 40°, 45°, 60°, or 70° was selected based on the strongest observed diffracted spherical waveform produced. Sizing calibration standards with 20%, 40%, 60%, and 80% electrical discharge machined (EDM) notches were used to accurately adjust screen range for the thickness and angles used. The difference between the peaks of the reflected radio frequency (RF) signal and the diffracted RF signals were calculated as a percentage of full screen width and used to calculate flaw height. Flaw characterization was performed in accordance with AWS D1.1 (AWS 2020c), Annex O. Tensile/cyclic stress acceptance criteria from AWS D1.1 and AWS D1.5 (AWS 2020a) were applied, and all flaws were considered rejectable. This team member also used TOFD to detect and size flaws. However, TOFD could not be used to size flaws in specimens with corner joints or T-joints. TOFD was used to size flaws in ERDC specimens 002, 003, 005, and 006.

Testing was done to verify the locations and dimensions of each flaw, confirm that each flaw was rejectable per AWS D1.5 (AWS 2020a) standards,

and ensure that the specimens contained no unintended discontinuities. The technician identified the start and stop locations of each flaw by locating the largest signal in three-dimensional space. This information was reported along with estimates of flaw length and height. The results of the verification testing confirmed that all flaws were as designed and were rejectable to AWS D1.5 standards for both compression/tension and static/cyclic members. The technician missed two flaws during blind verification testing, but the location and dimensions of these flaws were subsequently verified. The technician made no false calls during the verification testing. Laminations were not included in the summary of verification test results that follows because the technician did not report a flaw length for laminations.

The average absolute difference between the length reported by the manufacturer and the length estimated by the technician was 0.077 in. Similarly, the average absolute difference in reported height and estimated height was 0.024 in. These correspond to an average percent error in length of 8.0% and an average percent error in height of 15.3%. The maximum absolute difference in length was 0.49 in. (undersized), and the maximum absolute difference in height was 0.09 in. (oversized). Figure 13 and Figure 14 compare the reported and estimated lengths and heights. The diagonal line represents perfect agreement between reported and estimated length and height. Points located below the line of perfect agreement represent flaws that were undersized, and those above the line represent flaws that were oversized. The vertical distance from the line of perfect agreement indicates the difference between the measured and reported length.

If the flaw lengths and heights reported by the manufacturer are interpreted as the actual flaw length and height, then uncertainty in the technician's estimates of flaw length and height obtained using PE can be summarized using 90% confidence bounds based on lognormal distributions fit to the ratio of estimated to actual length or height (Chapter 4, Section 4.7). For length, the lower and upper confidence bounds were 0.67 and 1.61, respectively. For height, the lower and upper confidence bounds were 0.71 and 1.41, respectively.

Figure 13. Estimated and actual flaw length from verification testing.

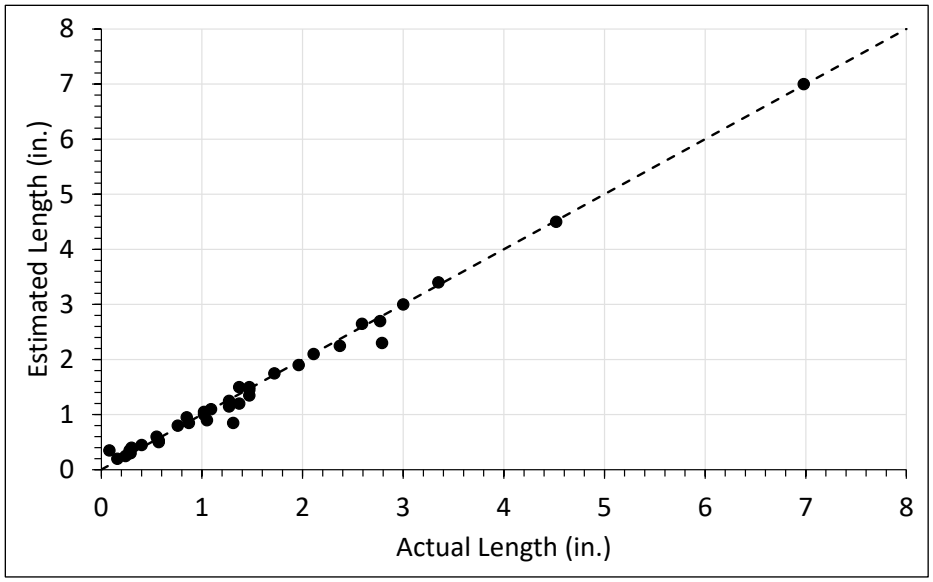
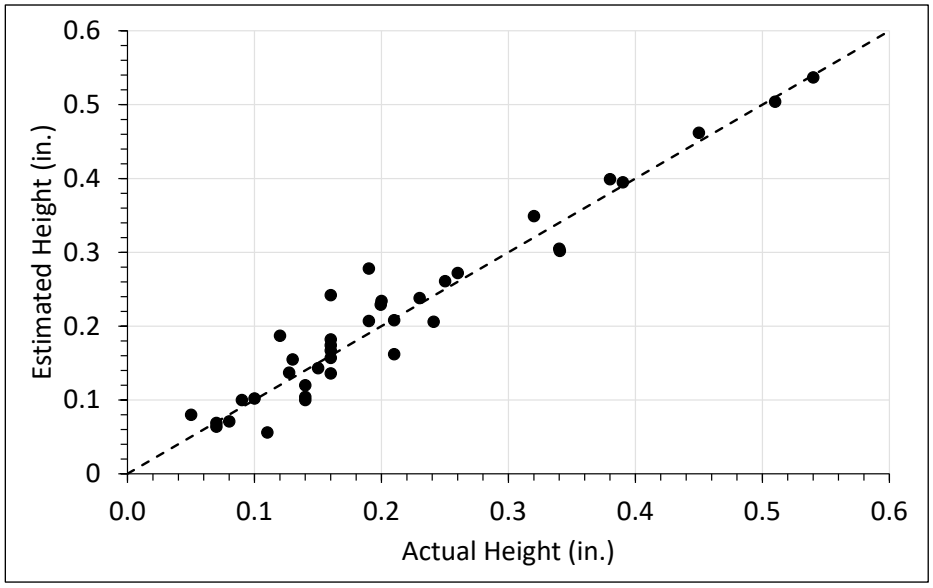


Figure 14. Estimated and actual flaw height from verification testing.



The Level III technician also verified the manufacturer's assignments of flaw category and subcategory. Overall, assignments of flaw category matched 92% of the time. Of the 30 flaws the manufacturer reported to be planar, the technician and the manufacturer agreed 29 times (96.7%). Of the eight flaws the manufacturer reported to be volumetric, the technician and the manufacturer agreed six times (80%). In terms of classification by subcategory, the technician and manufacturer agreed on 90% of the CRKs, 60% of the LOF-type flaws, 29% of the LOP-type flaws, 60% of the SLAG inclusions, and 67% of the POR type flaws.

3 Develop and Validate Ultrasonic NDT Procedures

The research team worked with individuals from academia and industry to investigate the capabilities and limitations of three advanced ultrasonic inspection techniques—PAUT, TFM/FMC, and TOFD—and to develop inspection procedures for the test specimens and round-robin testing. The team contracted with Michael Baker International (W912BU-18-D-0007) to complete this phase of the project. This work was accomplished through a 1.5-day workshop with a panel of experts from industry and academia followed by a week of inspection trials with Level III NDT technicians.

3.1 Expert Workshop on UT

The workshop was conducted virtually on 23 November and 24 November 2020. The expert panel consisted of Francesco Russo (Michael Baker International), Thomas Hay (TechKnowServ Corp), Ronnie Medlock (High Structures Steel), Jordan Wind (Bureau Veritas), Parrish Furr (Loenbro), Robert Connor (Purdue University), and Karl Frank (University of Texas at Austin). Russell Kok (Federal Highway Administration [FHWA]), Hoda Azari (FHWA), and Ray Momsen (Bureau Veritas) also participated at the invitation of the research team. The purpose of the workshop was to discuss the essential parameters that should be considered when developing and executing the inspection procedures. The workshop also included discussion on calibration methods, equipment selection, scanning techniques, data recording and interpretation, specimen design, scanning access and locations, use of UT modeling software (e.g., CIVA), testing of painted members, and flaw sizing techniques. Appendix C provides minutes from the workshop that were prepared by Michael Baker International. Following the workshop, two members of the expert panel developed an initial set of NDT procedures and scan plans for each specimen and NDT technique (i.e., PAUT, TFM/FMC, and TOFD). The research team reviewed and commented on the NDT procedures and scan plans, and a final version was accepted in July 2021.

3.2 Testing, Refinement, and Validation of Initial Procedures

The procedures and scan plans were tested, refined, and validated between 2 August 2021 and 6 August 2021 at ERDC in Vicksburg, Mississippi. Four ASNT Level III certified NDT technicians (Table 4) participated in the

inspection trials, applying the procedures and scan plans to detect and size flaws in the 12 ERDC specimens. An individual with Level III certification is assumed to be capable of developing, qualifying, and approving NDT procedures. This certification was required to ensure proficiency and to reduce the variability in technician skill. This enabled the research team to focus on assessing the capabilities of the ultrasonic testing techniques and procedures. Only one of the four technicians had previous experience with HSS inspection.

Table 4. Roster of participating Level III nondestructive testing (NDT) technicians.

Participant Name	Employer	ASNT Identification (ID)
Dale Cheek	Bureau Veritas	184306
Jordan Wind	Bureau Veritas	155638
Parrish Furr	Loenbro	187868
Thomas Hay	TechKnowServ	107162

The Level III NDT technicians rotated through the various specimens using the three different inspection techniques. The research team witnessed the calibration of each instrument to ensure conformance with the procedure. The technicians recorded the location and type of any indication greater than the time-corrected gain (TCG) on inspection forms prepared and furnished by the research team. Digital UT instruments report the TCG, which is the compensation of gain as a function of time for difference in amplitude of reflections from equal reflectors at different sound travel distances (CNDE, n.d.). This eliminates the need for the technician to plot a distance amplitude curve showing the relationship between signal amplitude and equal-sized reflecting surfaces at various distances from the transducer. Additionally, the start and stop locations of each indication were marked on the specimen with removable ink. For specimens with complicated geometry, the coordinate system proved to be a source of confusion throughout the testing week. The raw data files from each scan were collected by the research team at the end of each day of testing. Handwritten inspection forms were transcribed into spreadsheets by the research team. Figure 15 contains pictures from the inspection trials.

Figure 15. Level III technicians during inspection trials to test and validate procedures.

Inspections were self-paced. No limits were placed on the amount of time a technician could spend on any one specimen, although the research team suggested that 60–90 min might be needed for each specimen. Exact start and stop times were not recorded, but most inspections required 60–90 min. More complicated specimens, such as ERDC 002 or ERDC 007, required more than 90 min. Inspection rates seemed to increase as technicians became more familiar with the specimens and procedures. The technicians furnished all inspection equipment. They were asked to bring the equipment specified in the inspection procedures and any other equipment they typically use. Preliminary testing performed by the research team prior to the Level III field testing suggested that TOFD was not practical for several of the specimens due to joint geometry. Therefore, the majority of the inspections were performed with PAUT and TFM/FMC.

The week-long program to validate the procedures can be divided into three phases characterized by the amount of feedback provided to the technicians and the degree of collaboration between and among the research team and technicians. The initial phase of testing (phase A) was performed on Monday (2 August) and Tuesday morning (3 August). During the initial phase, the technicians were not provided with feedback on their performance, and the technicians were asked not to deviate from the previously developed inspection procedures and scan plans. During the

second phase of testing (phase B), the technicians were given general feedback and allowed to deviate from the scan plan to manually detect transverse indications. This phase of testing extended from Tuesday afternoon (3 August) to Wednesday morning (4 August). On Wednesday afternoon (4 August), the research team provided the technicians with feedback on their performance with respect to sizing flaws. Technicians were asked to brainstorm changes in the procedures that might improve performance. The final phase of testing (phase C) was completed with expanded inspection procedures that included manual rastering, line scans from more index offsets, focused and unfocused scans, scanning from all faces, and the use of smaller probes. The final phase of testing began on Wednesday afternoon (4 August) and finished on Friday afternoon (6 August).

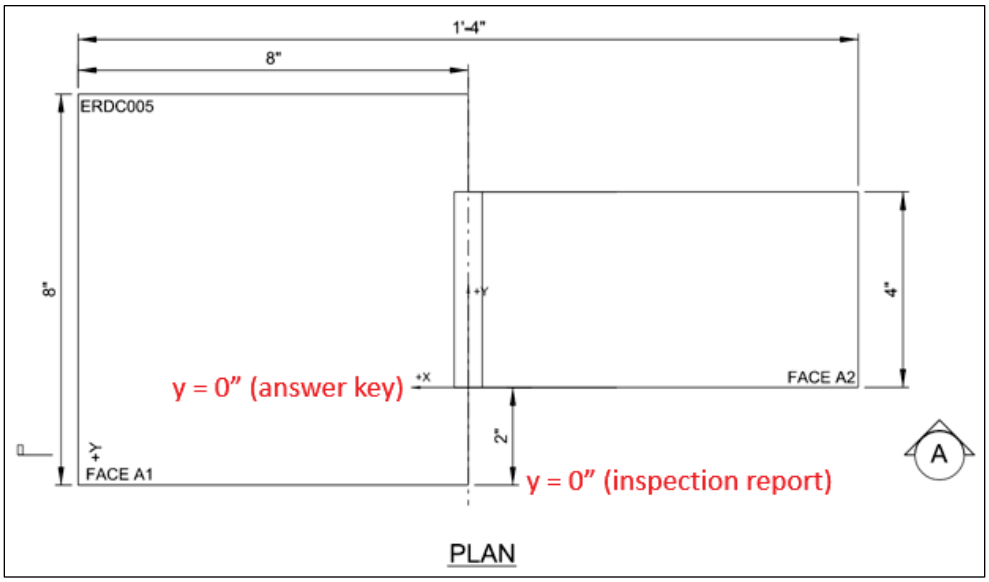
In total, 67 inspections were performed. This included 24 TFM/FMC inspections, 38 PAUT inspections, and 3 TOFD inspections. Two inspections were performed with manual contact PE UT, but those results were not considered in the analysis. Inspection rates varied by technician, so the number of inspections performed by each technician varied. Similarly, the number of inspections performed on each specimen varied. Table 5 shows the breakdown of inspections per specimen and per inspection technique. The letters A, B, and C in this table refer to the phase of testing. For example, technician 1 tested specimen 1 using PAUT during phases A and B.

Table 5. Summary of Level III inspection trials showing in what phase of testing (A, B, or C) each specimen was scanned by each technician.

Level III Technician	Technique	ERDC Specimen Number												TOTAL
		1	2	3	4	5	6	7	8	9	10	11	12	
1	PAUT	A, B	A		A, C	A, B	A	A	A	A, C	C		C	14
	TOFD													0
	TFM/FMC						C			C				2
2	PAUT			A	A	A	A	A, B	A			A		8
	TOFD		A			A						A		3
	TFM/FMC													0
	Manual UT	C				A								2
3	PAUT	B		C	C	C				C		C	C	8
	TOFD													0
	TFM/FMC	C	A			A	A	A	A	A	A	A	A, C	10
4	PAUT					A	A, C	C		C	C	B	C	8
	TOFD													0
	TFM/FMC	A	A	A	A	A	A	A	A	A	A	A	A	12
TOTAL		6	4	3	5	9	7	6	4	7	4	6	6	67

Data were manually transferred from the technicians' reports to a spreadsheet for evaluation. A spreadsheet was developed for each specimen prior to the round-robin to automatically evaluate the indications. The answer key for each specimen was prepared based on the as-built drawings provided by the manufacturer. In some cases, the research team was required to adjust the technician's inputs to match the coordinate system used in the spreadsheet. For instance, for specimen ERDC 005, shown in Figure 16, some of the technicians assumed that y = 0 was located at the edge of the wider plate, and so the y-coordinates recorded for each indication were adjusted by 2 in. to match the true coordinate system.

Figure 16. Common issue with the coordinate system on US Army Engineer Research and Development Center (ERDC) 005.



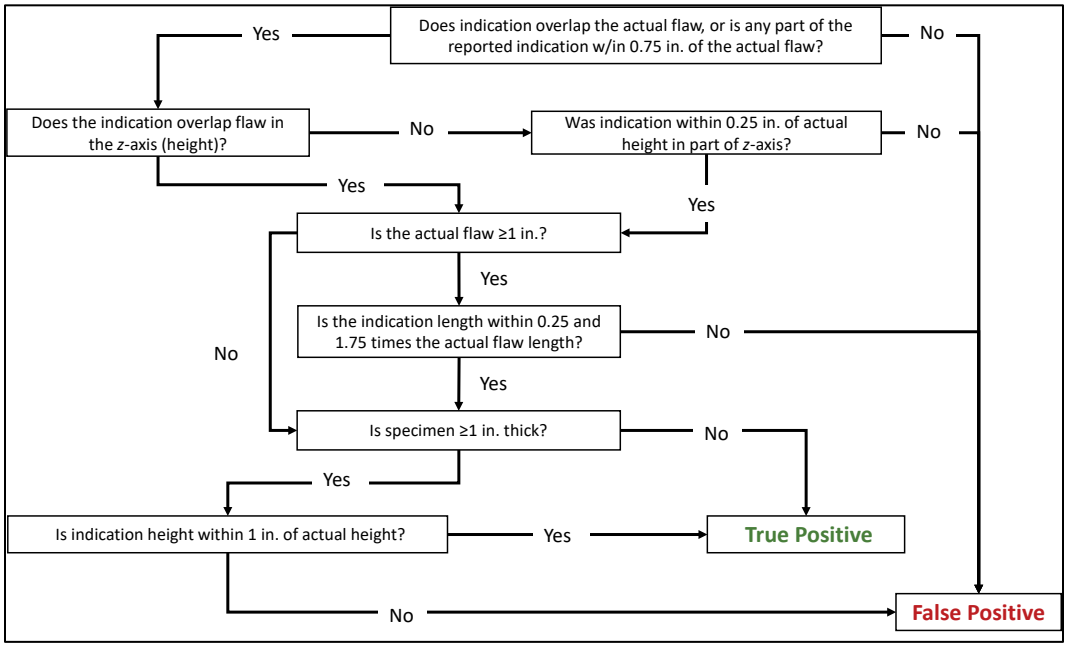
A set of criteria was established to distinguish between TP indications (i.e. hits or detections) and FP indications (false calls). Figure 17 outlines this system for distinguishing between hits and false calls. An indication was classified as a detection if it satisfied the following criteria and as a false call otherwise:

- The reported indication overlaps with the true flaw in the y-direction OR the beginning or end of the reported indication is within 0.75 in. of the beginning or end of the true flaw.
- The reported indication overlaps with the true flaw in the z-direction OR the beginning or end of the reported indication is within 0.25× specimen thickness of the beginning or end of the reported indication.

- For flaws greater than 1 in. in length, the reported length is between 0.25× the actual flaw length and 1.75× the actual flaw length.
- For specimens more than 1 in. thick, the reported height is within 1 in. of the actual flaw height.

These criteria, which are based loosely on AWS D1.8 (AWS 2016), reflect a tradeoff between rejecting valid data and accepting false data. In establishing these criteria, it was necessary to find a balance between criteria that were too stringent and criteria that were too liberal. If too stringent, this would lead to rejecting indications of actual flaws, thereby limiting the number of valid flaw size estimates. If too liberal, this would lead to accepting false indications and thereby introducing flaw size estimates that were not associated with a rejectable flaw.

Figure 17. Flowchart illustrating the system for distinguishing between hits and false calls.



Technicians were able to record indications in any order on the inspection report. To classify each indication as a hit or a miss, it was necessary to associate each indication with a single flaw, if possible. Each indication was compared to each flaw to determine if it satisfied the criteria for detection of that flaw. In some instances, multiple indications satisfied the criteria for a flaw. In these cases, the research team used their judgement to determine which indication best matched the true flaw. The research team did not use the technician's flaw type characterization in making this determination. This system of distinguishing between hits and misses was

considered generous. However, a generous threshold for associating indications with known flaws was necessary to obtain a sufficient number of flaw size comparisons for analysis. An example of the evaluation for one technician and one specimen is provided to illustrate the application of this system. Figures 18 and 19 illustrate the location of flaws and reported indications for a specimen. Tables 6 and 7 demonstrate the evaluation process that was used to compare the reported indications to the flaws and to determine if each indication was a hit or a false call. Ultimately, indication 1 was associated with flaw 1, indication 2 was associated with flaw 2, and indication 3 (Table 8) was classified as a false call.

Figure 18. Plan view of a specimen showing the location of flaws (*solid lines*) and reported indications (*dashed lines*) along the *y*-axis.

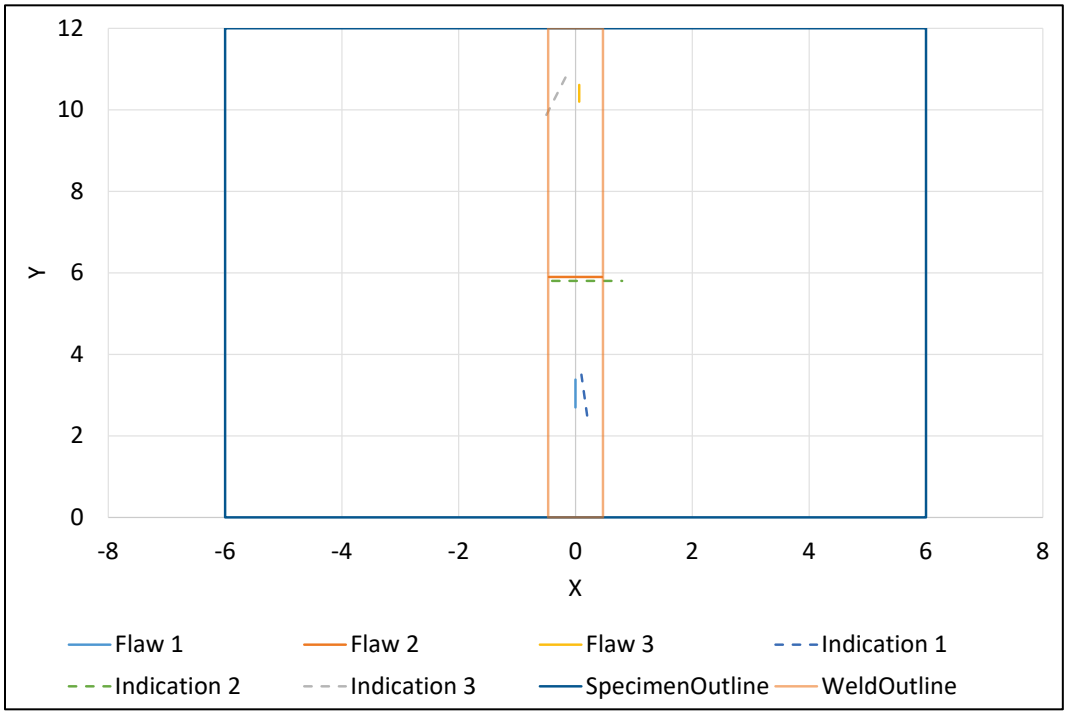


Figure 19. Section view of a specimen showing the location of flaws (*solid lines*) and reported indications (*dashed lines*) along the *z*-axis. Flaw 2 is a transverse flaw with a height of approximately 0.2 in. and a length of approximately 1 in. Opposite corners of the flaw are indicated in the figure.

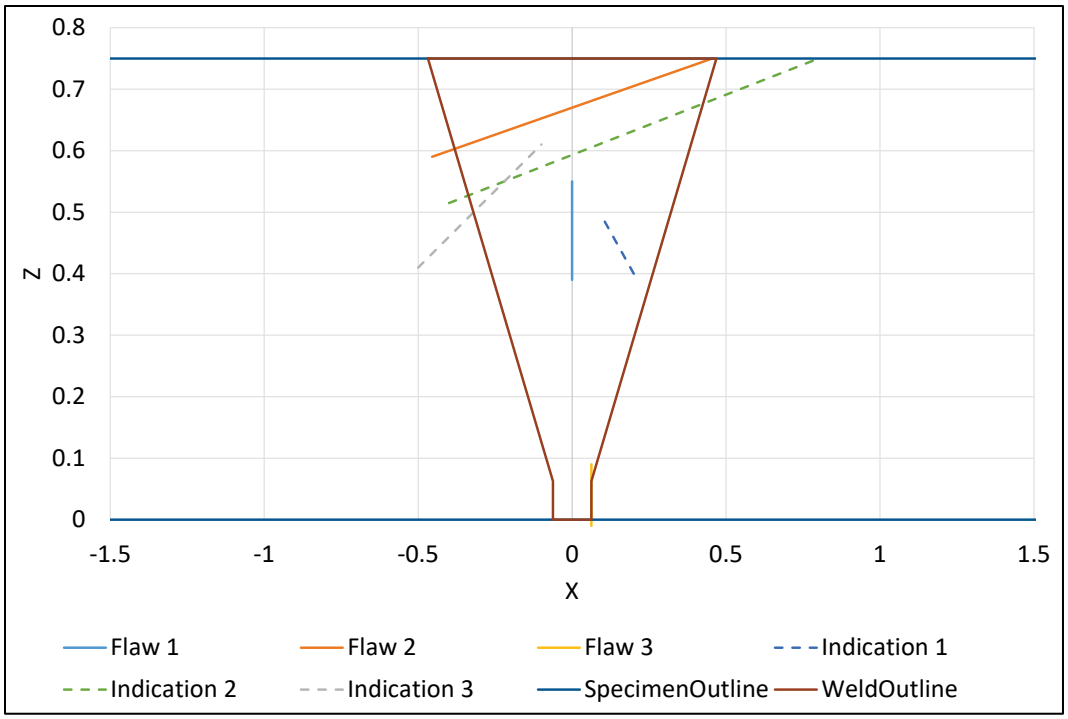


Table 6. Evaluation of indication 1.

Evaluation Criteria	Flaw 1	Flaw 2	Flaw 3
Does the reported indication overlap with the true flaw in the *y*-direction?	Yes	No	No
Is the reported start/stop within 0.75 in. of the actual start/stop?	N/A	No	No
Does the reported indication overlap with the true flaw in the *z*-direction?	Yes	No	No
Is the reported start/stop within 0.25*t* (0.1875 in.) of actual start/stop?	N/A	No	No
For flaws greater than 1 in. in length, is the reported length between 0.25× the actual flaw length and 1.75× the actual flaw length?	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)
For specimens more than 1 in. thick, is the reported height within 1 in. of the actual flaw height?	N/A (specimen thickness < 1 in.)	N/A (specimen thickness < 1 in.)	N/A (specimen thickness < 1 in.)
Result	Indication 1 = flaw 1		

Table 7. Evaluation of indication 2.

Evaluation Criteria	Flaw 1	Flaw 2	Flaw 3
Does the reported indication overlap with the true flaw in the *y*-direction?	No	Yes	No
Is the reported start/stop within 0.75 in. of the actual start/stop?	No	N/A	No
Does the reported indication overlap with the true flaw in the *z*-direction?	Yes	Yes	No
Is the reported start/stop within 0.25*t* (0.1875 in.) of actual start/stop?	N/A	N/A	No
For flaws greater than 1 in. in length, is the reported length between 0.25× the actual flaw length and 1.75× the actual flaw length?	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)
For specimens more than 1 in. thick, is the reported height within 1 in. of the actual flaw height?	N/A (specimen thickness <1 in.)	N/A (specimen thickness <1 in.)	N/A (specimen thickness <1 in.)
Result	Indication 2 = flaw 2		

Table 8. Evaluation of indication 3.

Evaluation Criteria	Flaw 1	Flaw 2	Flaw 3
Does the reported indication overlap with the true flaw in the *y*-direction?	No	No	Yes
Is the reported start/stop within 0.75 in. of the actual start/stop?	No	No	N/A
Does the reported indication overlap with the true flaw in the *z*-direction?	Yes	Yes	No
Is the reported start/stop within 0.25*t* (0.1875 in.) of actual start/stop?	N/A	N/A	No
For flaws greater than 1 in. in length, is the reported length between 0.25× the actual flaw length and 1.75× the actual flaw length?	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)	N/A (flaw length < 1 in.)
For specimens more than 1 in. thick, is the reported height within 1 in. of the actual flaw height?	N/A (specimen thickness <1 in.)	N/A (specimen thickness <1 in.)	N/A (specimen thickness <1 in.)
Result	Indication 3 = false call		

3.3 The Results of Procedure Validation

Test results from the Level III technicians were compiled and summarized for each specimen. To identify changes in performance as the inspection procedures were modified, the results from the initial and first round of testing (i.e., phases A and B) were compared to the final round of testing (i.e., phase C). The results from the initial and first round of testing were combined because limited changes were made in the B phase.

3.3.1 Flaw Detection

Table 9 summarizes the verification test results. The initial phases of testing (i.e., phases A and B) included 25 PAUT and 23 TFM/FMC inspections. The final phase of testing (i.e., phase C) included 17 PAUT and 3 TFM/FMC inspections. During the initial phases, technicians achieved a detection rate of 55% using PAUT and 67% using TFM/FMC. The technicians made 22 false calls during the PAUT inspections and 23 false calls during the TFM/FMC inspections. During the final phase, the detection rate for PAUT inspections increased to 83%, and the number of false calls decreased to 17. For TFM/FMC inspections, the detection rate increased to 92%, and the number of false calls decreased to 0. However, the sample size for the second round of testing with TFM/FMC was small. Only a single phase of TOFD testing was completed. The detection rate on three specimens was 70%, and no false calls were made. For those techniques for which there were multiple phases, the hit-to-call ratios calculated for the final phase of testing were greater than the hit-to-call ratios calculated from the initial round of testing, indicating improved efficiency.

Table 9. Summary of verification test results.

NDT Technique	Phase	Detection Rate	Number of Possible Hits	Number of False Calls	Hit-to-Call Ratio	Number of Inspections
PAUT	A, B	55%	80	22	67%	25
PAUT	C	83%	59	17	74%	17
TFM/FMC	A, B	67%	75	23	68%	23
TFM/FMC	C	92%	12	0	100%	3
TOFD	A	70%	10	0	100%	3

For all joint types, the detection rate improved from the initial to the final phase (Table 10). The number of false calls decreased from the initial phase of testing to the final phase in all but one case (Table 11). The number of false calls increased from 1 to 2 during the inspection of corner joints using PAUT. More false calls were made on the butt joints and T-joints when using PAUT than when using TFM/FMC. The hit-to-call ratio increased from the initial phase of testing to the final phase in all but one case (Table 12). The hit-to-call ratio decreased from 67% to 64% for butt joints inspected using PAUT.

Eight of the 12 specimens were inspected with PAUT during both the initial and final phases of testing. Table 13 summarizes the differences in performance between the phases. From the initial to the final phase of testing,

the detection rate increased on six of the specimens, decreased on one specimen, and remained the same on one specimen. The number of false calls remained relatively constant, with more false calls during the initial phase on four of the specimens.

Table 10. Detection rates by joint type.

Joint Type	Specimen Number	PAUT		TFM/FMC		TOFD
		A, B	C	A, B	C	
Butt joint	1–7	53% (66)	81% (26)	70% (43)	100% (4)	86% (7)
T-joint	8–10	63% (8)	89% (18)	61% (18)	75% (4)	—
Corner joint	11–12	67% (6)	80% (15)	64% (14)	100% (4)	33% (3)

Table 11. False calls by joint type.

Joint Type	Specimen Number	PAUT		TFM/FMC		TOFD
		A, B	C	A, B	C	
Butt joint	1–7	17	12	16	0	0
T-joint	8–10	4	3	4	0	—
Corner joint	11–12	1	2	3	0	0

Table 12. Hit-to-call ratio by joint type.

Joint Type	Specimen Number	PAUT		TFM/FMC		TOFD
		A, B	C	A, B	C	
Butt joint	1–7	67%	64%	65%	100%	100%
T-joint	8–10	56%	84%	73%	100%	—
Corner joint	11–12	80%	86%	75%	100%	100%

Table 13. Detection rate (total number of possible hits in parentheses) by specimen.

Specimen ID	PAUT		TFM/FMC	
	A, B	C	A, B	C
ERDC 001	50% (12)	—	75% (4)	100% (4)
ERDC 002	50% (4)	—	62.5% (8)	—
ERDC 003	50% (4)	75% (4)	100% (4)	—
ERDC 004	80% (10)	75% (8)	60% (5)	—
ERDC 005	58% (12)	100% (3)	83% (6)	—
ERDC 006	33% (9)	67% (6)	67% (6)	—
ERDC 007A	33% (9)	100% (3)	67% (6)	—
ERDC 007B	67% (6)	100% (2)	50% (4)	—
ERDC 008	50% (4)	—	75% (4)	—
ERDC 009	75% (4)	92% (12)	75% (8)	75% (4)
ERDC 010	—	83% (6)	33% (6)	—
ERDC 011	67% (6)	67% (3)	67% (6)	—
ERDC 012	—	83% (12)	62.5% (8)	100% (4)

Three of the 12 specimens were inspected with TFM/FMC during both the initial and final phases of testing. The detection rate increased on two of the specimens and remained the same on the third specimen. The total number of false calls was reduced from 23 to 0, although the number of inspections was also reduced (Table 14). The hit-to-call ratio increased from 68% during the initial phase to 100% during the final phase. Note that specimen 5 includes an unintended flaw between y = 3.5 in. and y = 4 in. While there is a real flaw at this location, the amplitude is less than the TCG. Therefore, it was considered an FP indication if the technicians recorded it as a flaw during this testing program.

Table 14. False calls by specimen.

	PAUT		TFM/FMC	
Specimen ID	A, B	C	A, B	C
ERDC 001	1	—	1	0
ERDC 002	0	—	0	—
ERDC 003	1	0	1	—
ERDC 004	1	2	2	—
ERDC 005	7	2	7	—
ERDC 006	3	5	4	—
ERDC 007A	0	2	1	—
ERDC 007B	4	1	0	—
ERDC 008	3	—	1	—
ERDC 009	1	1	0	0
ERDC 010	—	2	3	—
ERDC 011	1	0	2	—
ERDC 012	—	2	1	0

During the initial phases of testing using PAUT, the detection rate for planar defects (i.e., CRK, LOF, and LOP) was 52%, and the detection rate for volumetric defects (i.e., SLAG and POR) was 50%. The detection rate increased during the final phase of testing to 86% and 79%, respectively (Table 15). The detection rate for laminar defects decreased during the second phase of testing. However, this result was based on a single observation. During the initial phases of testing using TFM/FMC, the detection rate for planar defects was 67%, and the detection rate for volumetric defects was 79%. The detection rate increased during the final phase of testing to 89% and 100%, respectively. The detection rate for laminar defects increased during the final phase of testing, although the small number of observations during the final phase of testing may have biased these results.

Table 15. Detection rate (total number of possible hits in parentheses) by flaw type.

Flaw Shape	PAUT		TFM/FMC	
	A, B	C	A, B	C
Planar	52% (63)	86% (43)	67% (55)	89% (9)
Volumetric	50% (10)	79% (14)	79% (14)	100% (2)
Laminar	86% (7)	50% (2)	33% (6)	100% (1)

Although the objective of Level III testing was to evaluate the capability of the inspection technique and equipment, not the technician, it was not possible to completely eliminate the influence of the technician from performance results. For all technicians, the detection rate increased from the initial phases of testing to the final phase of testing (Table 16), although an exact comparison is difficult because both the number of inspections and the number of specimens inspected varied. Technicians also became more familiar with the specimens as the week progressed. Two of the technicians made more false calls during the final phase of testing than in the initial phases (Table 17). No clear trends were observed in the change in hit-to-call ratios between the initial phases of testing and the final phase (Table 18). For some technicians, the ratio increased, indicating greater efficiency, but for others, the ratio remained approximately the same or decreased. Again, this was likely because the number of inspections and specific specimens inspected varied between the testing phases.

Table 16. Detection rate (total number of possible hits in parentheses) by technician.

Technician	PAUT		TFM/FMC		TOFD
	A, B	C	A, B	C	
1	49% (37)	61% (18)	—	75% (4)	—
2	60% (30)	—	—	—	70% (10)
3	50% (4)	86% (22)	68% (31)	100% (8)	—
4	67% (9)	100% (19)	66% (44)	—	—

Table 17. False calls by technician.

Technician	PAUT		TFM/FMC		TOFD
	A, B	C	A, B	C	
1	14	9	—	0	—
2	6	—	—	—	0
3	0	3	7	0	—
4	2	5	16	—	—

Table 18. Hit-to-call ratio by technician.

Technician	PAUT		TFM/FMC		TOFD
	A, B	C	A, B	C	
1	56%	55%	—	100%	—
2	75%	—	—	—	100%
3	100%	86%	75%	100%	—
4	75%	79%	64%	—	—

The shortest flaw in the specimens was 0.08 in., and the longest flaw was 6.98 in. Flaws with lengths between 1 and 3 in. were detected more frequently than flaws with lengths that were less than 1 in. or longer than 3 in. (Table 19). The detection rate for the shortest flaw was 20% (1 of 5), and the detection rate for the longest flaw was 33% (2 of 6). Although POD was expected to increase with flaw length, some of the technicians explained that they were not expecting the specimens to possess long flaws, and so they attributed the signal variation to joint geometry rather than a flaw. For all flaw lengths, the detection rate was higher during the final phase of testing than during the initial phases of testing, with the exception of flaws with lengths greater than 4 in. during the second phase of TFM/FMC. The small number of observations during the final phase of testing may have biased the results.

Table 19. Detection rate (total number of possible hits in parentheses) by flaw length.

Flaw Length	Number of Flaws	PAUT		TFM/FMC	
		A, B	C	A, B	C
<0.5 in.	11	50% (16)	65% (20)	45% (20)	100% (3)
0.5–1 in.	8	47% (15)	100% (13)	64% (14)	100% (4)
1–1.5 in.	14	57% (28)	92% (13)	83% (23)	100% (4)
1.5–2 in.	2	75% (4)	100% (1)	75% (4)	—
2–3 in.	5	67% (9)	83% (6)	86% (7)	—
3–4 in.	2	50% (4)	100% (2)	100% (3)	—
>4 in.	2	50% (4)	75% (4)	25% (4)	0% (1)

Additional analysis would be needed to clarify which procedure changes had the greatest influence on inspection performance, but it is likely that clarification of the coordinate system and the implementation of manual raster scanning were at least partially responsible for the improvements in detection rate from phases A and B to phase C. Traditional line scans, even with phased array probes, are most effective at detecting flaws that are oriented along the longitudinal axis of the weld (perpendicular to the probe).

Therefore, introducing manual raster scans is likely to improve the detection of flaws that are skewed or transverse to the weld axis. Three transverse flaws (14% detection rate) were detected during the initial phases of testing (PAUT and TFM/FMC combined). Three transverse flaws (60% detection rate) were found during the final phase of testing (PAUT and TFM/FMC combined). Because the simplest way to detect a transverse flaw is by running the probe along the top of the weld, transverse flaws are still difficult to find with the larger probes typically used for PAUT and TFM/FMC and in welds that have not been ground smooth. The coordinate system was most confusing to communicate and visualize in skewed angle joints (i.e., specimens 10 and 12). For specimens 10 and 12, an increase in detection rate was observed between the initial and final phases of testing. In future research and real-world inspections, it is critical to ensure that the coordinate system is clearly labeled and communicated to the technician so that the findings can be interpreted correctly. Physically marking the indications on the specimen is one way to reduce confusion.

3.3.2 Flaw Sizing

The lengths and heights reported in as-built drawings were taken to be the actual lengths and heights of the specimens. The length reported by the technician (i.e., the measured length) was compared to the actual length to evaluate accuracy in length sizing, and the measured height was compared to the actual height to evaluate accuracy in height sizing. The error is the measured size minus the actual size, so a positive value indicates that the flaw was oversized (i.e., measured size greater than actual size), and a negative value indicates that the flaw was undersized (i.e., measured size less than the actual size). The absolute error was calculated as the absolute value of the measured length minus the actual length. The percent error was calculated as the error divided by the actual size, and the absolute percent error was calculated as the absolute error divided by the actual size.

3.3.2.1 Flaw Length

When technicians used PAUT, they tended to underestimate flaw length in the initial phases of testing, but they displayed no tendency to under- or overestimate length in the final phase of testing (Table 20). The average measurement errors were –0.1 in. and 0 in., and the average absolute measurement errors were 0.47 in. and 0.21 in. in the initial phases and final phase, respectively. The percent absolute error decreased from the initial phases to the final phase of testing, indicating improved accuracy in

length sizing. When technicians used TFM/FMC, they tended to underestimate flaw length in the initial phases of testing and overestimate length in the final phase of testing. The average measurement errors were –0.03 in. and 0.13 in., and the average absolute measurement errors were 0.27 in. and 0.15 in. in the initial phases and final phase, respectively. The percent absolute error decreased from the initial phases to the final phase of testing, indicating improved accuracy in length sizing. Using TOFD, the technicians tended to overestimate the flaw length. The improvement in length sizing accuracy from the initial phases of testing to the final phase of testing was greater in PAUT than in TFM/FMC. However, the accuracy in the initial phases was greater with TFM/FMC; therefore, improvement was harder to achieve.

Table 20. Error analysis for flaw length estimates by NDT technique.

NDT Technique	Statistic	Testing Phase			
		A, B		C	
		Error	Absolute Error	Error	Absolute Error
PAUT	Min./max. (in.)	–5.08/1.28	NA	–1.3/0.47	NA
	Average (in.)	–0.10	0.47	0.00	0.21
	St. dev. (in.)	0.91	0.78	0.31	0.22
	Average (%)	27%	52%	18%	28%
	Sample size (n)	44		49	
TFM/FMC	Min./max. (in.)	–0.97/0.86	NA	–0.15/0.45	NA
	Average (in.)	–0.03	0.27	0.13	0.15
	St. dev. (in.)	0.36	0.24	0.14	0.11
	Average (%)	16%	35%	27%	30%
	Sample size (n)	50		11	
TOFD	Min./max. (in.)	–0.35/0.07	NA	—	—
	Average (in.)	0.07	0.18	—	—
	St. dev. (in.)	0.21	0.13	—	—
	Average (%)	27%	39%	—	—
	Sample size (n)	7		—	

To provide a visual representation of the ability to accurately estimate flaw lengths, actual and estimated flaw lengths were plotted for PAUT and TFM/FMC (Figures 20 and 21). The dashed diagonal line represents the line of perfect agreement between the actual and measured length. Flaws that plotted below the 1:1 line were undersized, and flaws that plotted above the 1:1 line were oversized. The distance from the 1:1 line indicates the error in the measured length. For flaws detected with PAUT, the graph

illustrates that the error decreased from phase A to B to C. The root-mean-square error (RMSE) describes the average difference between estimated and actual length. The RMSE for length estimates was 1.034 in. during phase A, 0.33 in. during phase B, and 0.311 in. during phase C. The correlation describes the linear association between the actual and estimated flaw length. The correlation increased from 0.68 (phase A) to 0.89 (phase B) to 0.97 (phase C; Figure 20). The very low correlation during phase A can mostly be attributed to a single observation with a large measurement error. If this outlier is removed, the correlation during phase A increases to 0.8459. Although not as dramatic, improvements in the correlation between measured and actual flaw length can still be seen even after this outlier has been removed. For flaws detected with TFM/FMC, Figure 21 shows that flaws detected during phase C tended to be oversized. The RMSE for TFM/FMC length estimates was 0.357 in. during phase A and 0.184 in. during phase C. The correlation between the actual flaw length and the measured flaw length was 0.93 for both phases A and C.

Figure 20. Comparison of estimated and actual flaw length using PAUT.

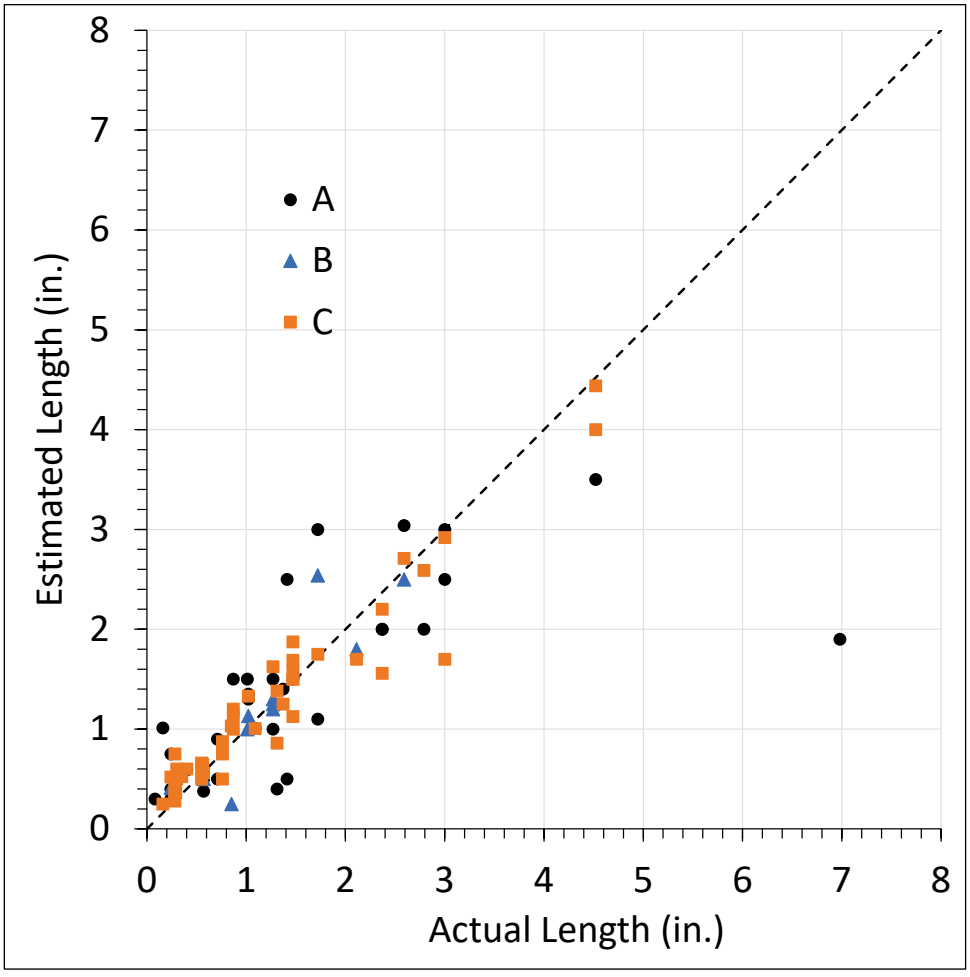
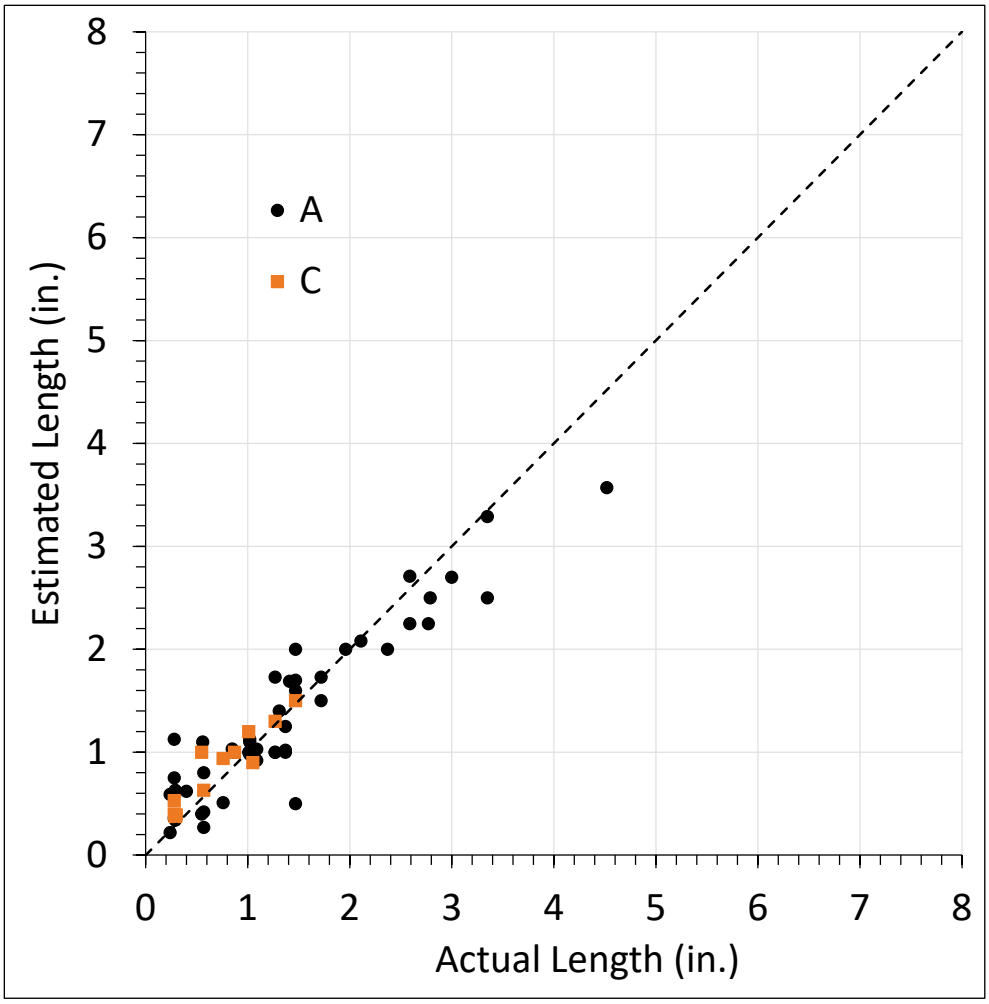


Figure 21. Comparison of estimated and actual flaw length using TFM/FMC.



Length sizing accuracy was also compared across technicians, flaw types, joint types, and flaws (Table 21). Similar to the overall results, accuracy tended to improve from the initial phases to the final phase of testing. No patterns were observed in the accuracy across technicians or flaw types. Results from the final phase of testing using TFM/FMC and PAUT suggest that accuracy in length sizing is better in butt joints and worse in T- and corner joints.

Table 21. Error analysis for flaw length measurements by joint type.

Joint Type	Statistic	PAUT		TFM/FMC	
		Testing Phase C		Testing Phase C	
		Error	Absolute Error	Error	Absolute Error
Butt joints	Min./max. (in.)	−1.3/0.355	NA	−0.15/0.19	NA
	Average (in.)	−0.11	0.26	0.03	0.11
	St. dev. (in.)	0.38	0.29	0.12	0.06
	Average (%)	9%	25%	4%	11%
	Sample size (n)	21		4	
T-joints	Min./max. (in.)	−0.52/0.405	NA	0.03/0.45	NA
	Average (in.)	0.05	0.17	0.20	0.20
	St. dev. (in.)	0.22	0.15	0.18	0.18
	Average (%)	11%	17%	33%	33%
	Sample size (n)	16		3	
Corner joints	Min./max. (in.)	−0.26/0.47	NA	0.09/0.25	NA
	Average (in.)	0.14	0.18	0.16	0.16
	St. dev. (in.)	0.18	0.13	0.06	0.06
	Average (%)	41%	47%	46%	46%
	Sample size (n)	12		4	

3.3.2.2 Flaw Height

When technicians used PAUT, they tended to overestimate flaw height in the initial phases of testing and underestimate flaw height in the final phase of testing (Table 22). The average measurement errors were 0.02 in. and −0.04 in., and the average absolute measurement errors were 0.12 in. and 0.07 in. in the initial phases and final phase, respectively. The percent absolute error decreased from the initial phases to the final phase of testing, indicating improved accuracy in height sizing. When technicians used TFM/FMC, they tended to underestimate flaw height in the initial and final phases of testing. The average measurement errors were −0.03 in. and −0.02 in., and the average absolute measurement errors were 0.08 in. and 0.05 in. in the initial phases and final phase, respectively. The percent absolute error decreased from the initial phases to the final phase of testing, indicating improved accuracy in height sizing. When using TOFD, the technicians tended to underestimate flaw height. In both the initial phases of testing and the final phase, the average absolute height sizing error and percent error were smaller with TFM/FMC than with PAUT. The improvement in height sizing accuracy from the initial phases of testing to the final phase of testing was greater in PAUT than in TFM/FMC. However, the

accuracy in the initial phases was greater with TFM/FMC; therefore, improvement was harder to achieve.

Table 22. Error analysis for flaw height measurements.

NDT Technique	Statistic	Testing Phase			
		A, B		C	
		Error	Absolute Error	Error	Absolute Error
PAUT	Min./max. (in.)	−0.3/0.74	NA	−0.26/0.22	NA
	Average (in.)	0.02	0.12	−0.04	0.07
	St. dev. (in.)	0.17	0.12	0.09	0.06
	Average (%)	28%	64%	−9%	38%
	Sample size (n)	38		48	
TFM/FMC	Min./max. (in.)	−0.259/0.21	NA	−0.13/0.08225	NA
	Average (in.)	−0.03	0.08	−0.02	0.05
	St. dev. (in.)	0.10	0.07	0.06	0.04
	Average (%)	−5%	43%	−3%	27%
	Sample size (n)	48		10	
TOFD	Min./max. (in.)	−0.25/−0.07	NA	—	—
	Average (in.)	−0.07	0.09	—	—
	St. dev. (in.)	0.11	0.09	—	—
	Average (%)	−13%	25%	—	—
	Sample size (n)	6		—	

To provide a visual representation of height sizing accuracy, the actual flaw height was plotted against the measured flaw height for all flaws detected using PAUT and TFM/FMC (Figures 22 and 23). For flaws detected with PAUT, the graph illustrates that the error decreased from phases A and B to phase C. The RMSE for height estimates was 0.163 in. during phase A, 0.099 in. during phase B, and 0.093 in. during phase C. The correlation between the actual flaw height and the measured flaw height was 0.41 during phase A, decreased to 0.36 during phase B, and increased to 0.71 during phase C. As with length, there was a single outlier with a large measurement error during phase A. The correlation shows improvements in flaw measurement capabilities even when this outlier is removed. For flaws detected with TFM/FMC, the graph illustrates that flaws detected during A and C tended to be undersized. The RMSE was 0.1068 in. during phase A and 0.057 during phase C. The correlation between the actual and estimated flaw height increased from 0.71 (in A) to 0.78 (in C).

Figure 22. Comparison of estimated and actual flaw height using PAUT.

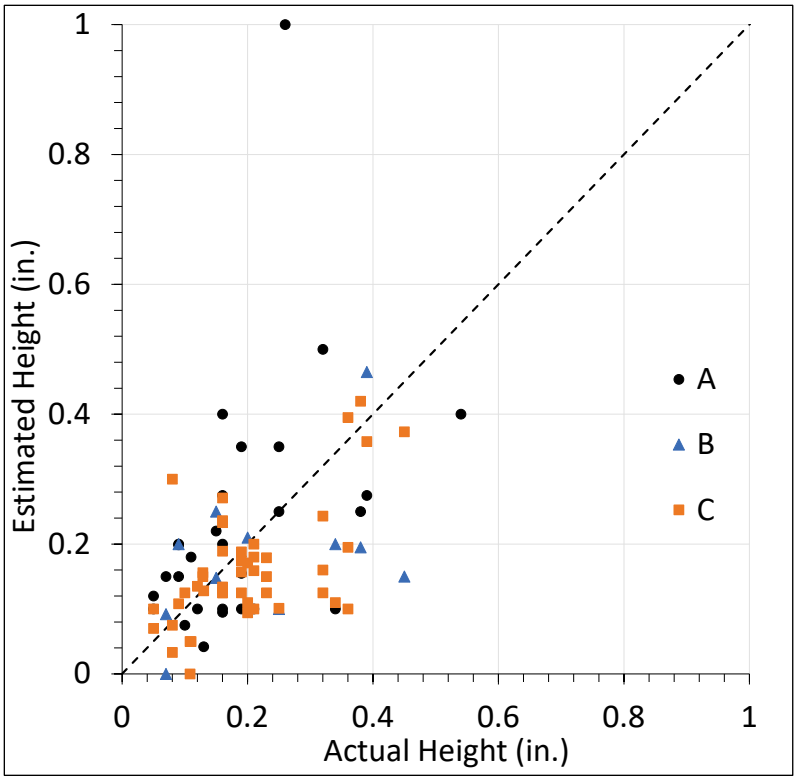
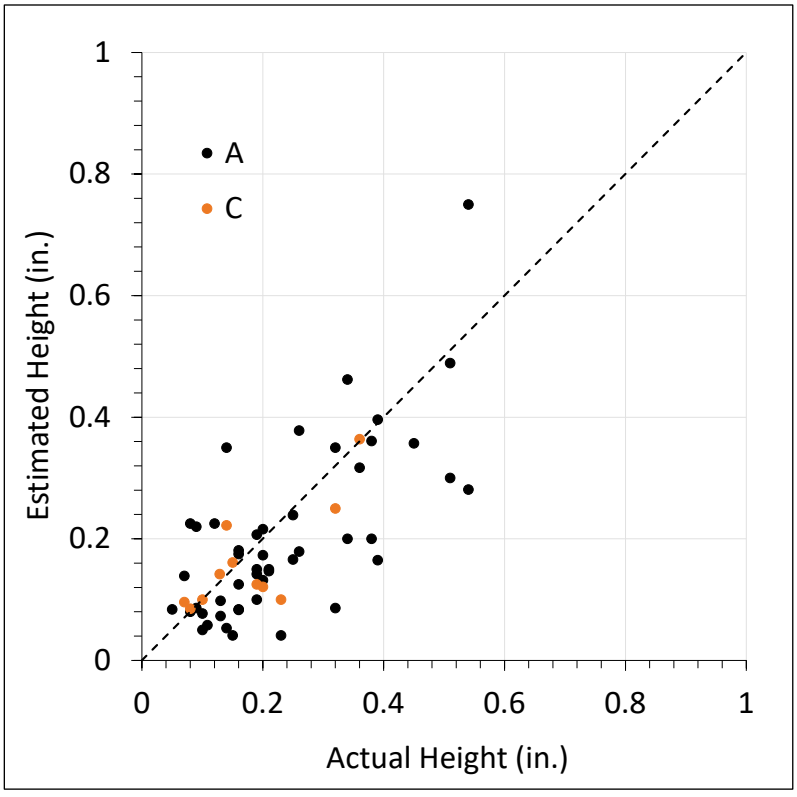


Figure 23. Comparison of estimated and actual flaw height using TFM/FMC.



Height sizing accuracy was also compared across technicians, flaw types, joint types, and flaws. Similar to the overall results, accuracy tended to improve from the initial phases to the final phase of testing, and accuracy tended to be better with TFM/FMC than with PAUT. No patterns were observed in the accuracy across joint types or flaw shapes. Using PAUT, the average absolute error and percent error in height sizing were larger for technician 1 than for the other technicians.

3.3.3 Flaw Characterization

In addition to detecting and sizing the flaws, the technicians were asked to characterize the flaw type (Table 23). During the initial phases of testing, the characterization accuracy was 43% for flaws detected with PAUT and 52% for flaws detected with TFM/FMC. During the final phase of testing, the classification accuracy increased to 55% for flaws detected with PAUT and 82% for flaws detected with TFM/FMC. The technician who performed the NDT tests using TOFD did not characterize flaw subcategories, so the accuracy of flaw subcategory is not reported in Table 23. Accuracy in flaw category characterization was above 70% for all phases of testing and all inspection techniques. A higher accuracy was achieved with TFM/FMC than with PAUT and TOFD.

Table 23. Accuracy in flaw characterization.

NDT Technique	Phase	Characterization Accuracy (%)		Sample Size
		Flaw Subcategory	Flaw Category	
PAUT	A, B	43%	84%	44
PAUT	C	55%	78%	49
TFM/FMC	A, B	52%	90%	50
TFM/FMC	C	82%	91%	11
TOFD	—	—	71%	7

It was easier for technicians to characterize flaw categories than flaw subcategories (Table 24). CRK, LOF, and LOP were correctly identified as planar flaws at least 80% of the time across all PAUT and TFM/FMC inspections. While POR was always correctly classified as a volumetric flaw, SLAG was often misclassified as a planar flaw. LAMs were almost always accurately identified, though one LAM was misclassified as LOF using PAUT.

Table 24. Percent accuracy of flaw characterization by NDT technique and testing phase.

Flaw Subcategory	Percent Accuracy (%) by NDT Technique and Testing Phase			
	PAUT		TFM/FMC	
	A, B	C	A, B	C
Cracks (CRK)	82%	96%	91%	100%
Lack of fusion (LOF)	100%	80%	100%	100%
Lack of penetration (LOP)	100%	100%	100%	NA
Porosity (POR)	100%	100%	100%	100%
Slag inclusion (SLAG)	25%	11%	50%	0%
Lamination (LAM)	100%	0%	100%	100%

Tables 25 and 26 report the ability of technicians to characterize flaw category. Planar flaws were correctly identified as planar at least 88% of the time across all PAUT and TFM/FMC inspections. Using PAUT, the technicians frequently misclassified volumetric flaws as planar. In the final phase of PAUT testing, volumetric flaws were correctly classified in 27% of the detections. Using TFM/FMC, technicians were more likely to correctly identify volumetric defects. The characterization accuracy was 73% in the initial phases of testing and 50% in the final phase of testing.

Table 25. Percent accuracy in flaw type characterization using PAUT.

Actual Flaw Type	Testing Phase and Reported Flaw Type (PAUT)					
	A, B			C		
	Planar	Volumetric	Laminar	Planar	Volumetric	Laminar
Planar	88%	12%	0%	95%	2.5%	2.5%
Volumetric	60%	40%	0%	73%	27%	0%
Laminar	0%	0%	100%	100%	0%	0%

Table 26. Percent accuracy in flaw type characterization using TFM/FMC.

Actual Flaw Type	Testing Phase and Reported Flaw Type (TFM/FMC)					
	A, B			C		
	Planar	Volumetric	Laminar	Planar	Volumetric	Laminar
Planar	95%	5%	0%	100%	0%	0
Volumetric	27%	73%	0%	50%	50%	0%
Laminar	0%	0%	100%	0%	0%	100%

After the inspection trials, the Level III technicians revised the inspection procedures for PAUT, TFM/FMC, and TOFD and submitted the final procedures to the research team. Final inspection procedures, prepared by the

research team, required manual rastering, line scans from multiple offsets, scanning from all faces, the use of smaller probes, and a transverse scan performed along the top of the weld cap. Appendix E includes the final inspection procedures.

4 Round-Robin Testing

The research team contracted with Michael Baker International to recruit Level II NDT technicians who were certified in accordance with Recommended Practice No. SNT-TC-1A (ASNT 2020b) to participate in round-robin experiments. Michael Baker International subcontracted with several NDT service providers to supply a pool of candidates who were willing and able to participate in round-robin experiments. The service providers selected the technicians who formed the pool of candidates. All technicians were paid for their participation as employees of the NDT service providers. Prior to the round-robin experiments, the research team administered a prequalification exam to this pool of candidates. The research team selected technicians from the pool of candidates to participate in round-robin testing based on their prequalification exam performance. The objective of the performance demonstration was to establish the capabilities of each NDT technician with respect to two of the three UT techniques (i.e., PAUT, TFM/FMC, or TOFD). However, the industry was unable to supply a suitable number of technicians experienced with TFM/FMC or TOFD. Therefore, the performance demonstration and subsequent round-robin trials focused primarily on PAUT.

4.1 Performance Qualification

The Level II performance qualification demonstration consisted of a written specific examination and a hands-on practical examination. A specific exam tests the technician's understanding of procedures, codes, specifications, and equipment or instrumentation for an NDT technique used by the employer (ASNT 2020a). A practical exam tests the ability of a technician to apply those techniques to the materials being tested. Prequalification events were held in Philadelphia, Pennsylvania; New Orleans, Louisiana; and Houston, Texas. Prior to prequalification, participants were provided with an orientation letter outlining the purpose of the research and the expectations for the participants. Each participant was asked to fill out a background form with information about their employment, training, certification, and experience and to sign a confidentiality agreement. On the day of prequalification, a member of the research team delivered a short presentation to all participants to summarize the inspection procedures and answer questions about the examinations. Participants were asked to work independently and were observed by the research team throughout the exam.

The specific exam consisted of 30 questions that evaluated the candidate's understanding of inspection procedures, codes, standards, specifications, and equipment. Candidates were provided with a formula sheet for reference and allowed two hours to complete the exam. The practical examination was a hands-on test intended to evaluate the candidate's ability to conduct UT in accordance with the furnished procedures. Candidates were allowed two hours to test two welded samples totaling 24 linear inches. Four of the AWS Bridge Kit specimens were used for prequalification. During the two-hour time period, the candidates were asked to calibrate their instruments, test both plates, and document the test findings. The intention was for each specimen to be tested with a different technique (i.e., PAUT, TFM/FMC, or TOFD); however, most of the candidates were only trained in one of these techniques. Therefore, a number of the candidates performed conventional UT (i.e., PE) on the second specimen. Participants were asked to furnish their own equipment for prequalification. All equipment was National Institute of Standards and Technology (NIST) calibrated. The research team provided hard copies of the inspection procedures and reporting forms.

Inspection results were evaluated based on detection rate, false call rate, and flaw size estimation. Because there are no established tolerances or criteria for judging whether or not a flaw size estimate is accurate, the requirements in AWS D1.8 (AWS 216), Annex F, were adapted and used with the detection criteria described in Chapter 3. The AWS D1.8 Seismic Supplement requires NDT technicians to pass a POD detection exam prior to all testing of demand-critical connections. For flaw length, the estimated length of a detected flaw was required to be within the interval (−0.5 in., +1 in.) relative to the known length of the flaw. For height, the estimated height of a detected flaw was required to be within the interval (−0.25 in., +0.25 in.) relative to the known height of the flaw. These tolerances were used to quickly assess flaw sizing accuracy across the pool of participants by comparing the difference between the number of flaws detected and the number of flaws that were detected and sized. These tolerances were not applied to the round-robin inspection results because the actual accuracy in flaw sizing was quantified. Candidates were ranked based on their performance on the specific and practical exams. In total, 27 technicians participated in the prequalification demonstrations, and 21 of the participants were invited to participate in the round-robin inspections (Table 27).

Table 27. Summary of participation in prequalification sessions.

Prequalification Location	Prequalification Date	Number of Technicians	Number of Technicians Selected
Philadelphia, PA	1 November 2022	5	3
New Orleans, LA	29–30 November 2022	6*	4*
Houston, TX	2–3 December 2022	16*	14*

*One technician attended in both New Orleans and Houston and is included in the New Orleans tally.

NDT technicians were selected from a pool of proposed candidates based on how well they performed on the practical and specific exams. The research team administered prequalification exams in Philadelphia, Pennsylvania, on 2 November 2021; in New Orleans, Louisiana, on 29–30 November 2021; and in Houston, Texas, on 2–3 December 2021. Table 28 summarizes the results from the performance qualification events. Candidates are grouped in the table based on the location and date of testing but are otherwise listed in no particular order. A unique candidate identification (ID) number was randomly assigned to each candidate. Table 28 lists the practical and written exam scores for each of the 27 technicians who participated in performance qualification events in Philadelphia, New Orleans, and Houston.

Two scores were calculated from each candidate's results on the practical exam (Table 28). D is the fraction of scanned flaws that were detected and sized within tolerance. No credit was given for flaws that were detected but not sized within tolerances. F is the ratio of false calls to total indications (i.e., the false discovery rate [FDR]). An overall score, R, was calculated from D and F as follows: $R = 0.5 \cdot (1 + D - F)$ Each candidate's performance on the written exam was rated in terms of the proportion of correct responses. However, analysis of the performance qualification results showed that performance on the written exam was not correlated with performance on the practical exam. Therefore, results on the written exam were not used as a basis for accepting technicians into the round-robin study.

Table 28. Performance qualification results.

Candidate Identifier	NDT Techniques		Practical Exam Scores			Specific Exam	Accepted	Technician Identifier
	Method 1	Method 2	*D*	*F*	*R*			
1	PAUT	TFM/FMC	0.67	0.38	0.65	0.73	True	4
2	PAUT	PAUT	0.67	0.00	0.83	0.60	True	1
3	PE	—	DNF*	DNF*	DNF*	0.53	False	—
4	PE	PE	0.50	0.25	0.63	0.73	False	—
5	PAUT	PE	0.83	0.00	0.92	0.87	True	2
6	PAUT	PE	0.67	0.17	0.75	0.43	True	—
7	PAUT	TFM/FMC	1.00	0.00	1.00	0.67	True	5
8	PAUT	PE	0.17	0.75	0.21	0.33	False	—
9	PAUT	PE	1.00	0.00	1.00	0.53	True	7
10	PAUT	PE	0.67	0.33	0.67	0.63	True	3
11	TOFD	TOFD	0.88	0.22	0.83	0.73	True	12
12	PAUT	PAUT	1.00	0.00	1.00	0.67	True	6
13	PAUT	PE	1.00	0.00	1.00	0.70	True	18
14	PAUT	TOFD	0.80	0.20	0.80	0.67	True	—
15	PAUT	PAUT	1.00	0.00	1.00	0.63	True	15
16	PAUT	PAUT	0.83	0.17	0.83	0.67	True	16
17	PAUT	PE	0.83	0.17	0.83	0.70	True	8
18	PAUT	PAUT	0.67	0.17	0.75	0.43	True	11
19	PAUT	TFM/FMC	0.00	1.00	0.00	0.37	False	—
20	PAUT	PAUT	0.83	0.17	0.83	0.73	True	14
21	PAUT	PE	1.00	0.00	1.00	0.80	True	17
22	PAUT	PAUT	1.00	0.00	1.00	0.70	True	9
23	PAUT	PAUT	0.33	0.71	0.31	0.60	False	—
24	PAUT	PAUT	0.50	0.17	0.67	0.80	True	10
25	PAUT	PAUT	0.50	0.17	0.67	0.67	True	13
26	PAUT	PE	0.00	1.00	0.00	0.33	False	—
27	PAUT	PAUT	0.67	0.20	0.73	0.70	True	—

*DNF = Did not finish.

Candidates were accepted into the study based strictly on their performance on the practical exam. Practical exam scores were clustered, perhaps because each candidate scanned only a small number of test specimens. Of the 27 technicians who participated in the prequalification testing, 21 were invited into the study based on their results. No hard cutoffs were established for acceptance into the study. However, the score of the candidate with the lowest practical exam score accepted into the study was $R = 0.65$. Three of the candidates had to drop out of the study between prequalification and round-robin testing because they changed employers.

For the 18 technicians who participated in the study, Table 29 reports the number of flaws scanned, the number of flaws detected, the number of flaws detected and sized within tolerances, and the number of false indications. This information is being provided to support the comparison of performance metrics from performance qualification with performance metrics from round-robin testing. Technician 12 scanned more plates during performance qualification than did other technicians because he would not comply with the instructions given by the research team, but this did not affect study results.

Table 29. Performance qualification results.

Technician Identifier	Flaws Scanned (*N*)	Flaws Detected			Total Missed (FN)	False Indications (FP)	Total Indications (PP)
		Flaws Sized		Total Detected (TP)			
		Within Tolerance (*S*)	Not within Tolerance (NS)				
1	6	4	0	4	2	0	4
2	6	5	1	6	0	0	6
3	6	4	0	4	2	2	6
4	6	4	1	5	1	3	8
5	6	5	0	5	1	1	6
6	6	6	0	6	0	0	6
7	6	6	0	6	0	0	6
8	6	5	0	5	1	1	6
9	6	6	0	6	0	0	6
10	6	3	2	5	1	1	6
11	6	4	1	5	1	1	6
12	16	14	0	14	2	4	18
13	6	3	2	5	1	1	6
14	6	5	0	5	1	1	6
15	6	6	0	6	0	0	6
16	6	5	0	5	1	1	6
17	6	6	0	6	0	0	6
18	6	6	0	6	0	0	6

Note: TP is true positives, FN is false negatives, FP is false positive, and PP is total indications.

4.2 Round-Robin Testing

The round-robin inspections were performed at ERDC over three one-week periods between 7 February 2022 and 25 March 2022. The objective of this phase of the research was to evaluate the capability of NDT

technicians inspecting common HSS joints using the inspection procedures developed by the research team based on the findings from the previous phase of research. Eighteen Level II NDT technicians certified in accordance with Recommended Practice No. SNT-TC-1A (ASNT 2020b) participated in the round-robin inspections. Level II technicians are certified by their employers and are assumed to be capable of calibrating the equipment, implementing the inspection procedure, and interpreting and evaluating inspection results. All 18 technicians had experience performing inspections for the oil and gas industry; six technicians had previous HSS inspection experience. Table 30 lists the NDT inspection firms that provided technicians and the number of technicians participating during each week of testing.

Table 30. Number and affiliation of Level II NDT technicians by round-robin testing week.

Company Name	Number of Technicians by Week			Total
	7–11 Feb. 2022	14–18 Feb. 2022	21–25 March 2022	
Applied Technical Services	—	—	4	4
Bureau Veritas	2.5*	5.5*	3	11
MISTRAS Group	1	—	—	1
TechKnowServ	1	—	1	2
Total	4.5	5.5	8	18

*One technician from Bureau Veritas distributed his effort across two one-week periods.

At the beginning of each week, the research team delivered a short presentation discussing the motivation for and objectives of the research and summarizing the inspection procedures. The research team was available throughout the week to answer questions. Technicians were instructed to work independently (Figure 24) and not to discuss the specimens or their work with other participants. However, the inspections were staggered over a five-day period, and the research team was unable to control communications between technicians during breaks or after hours. Therefore, the research team cannot be certain that the technicians did not discuss the specimens or their findings with other round-robin participants.

Technicians were asked to inspect 20 specimens (ERDC and AWS Bridge Kit) during the 40 hr work week. No time restrictions were placed on the technicians, although a time limit of 60–90 min per specimen was suggested. Although exact start and stop times were not recorded, the inspection rate varied by inspection and specimen. The order of the inspections varied for each technician and was determined in advance by the research

team. Progress was monitored throughout the week, and the schedule was adjusted as necessary. After completing each specimen, the technicians turned in their inspection report to the research team. The research team briefly reviewed the report, asked for clarification if needed, and then provided the next specimen. The research team copied the handwritten report into an Excel spreadsheet for evaluation in real time.

All of the inspections described in this report were completed using PAUT on instruments manufactured by the Olympus Corporation. Table 31 summarizes the inspection equipment. Technicians furnished their own inspection equipment. Technicians recorded the location and category of all indications greater than the TCG on standard inspection forms prepared and furnished by the research team. In addition to recording their findings in a table, the technicians were asked to sketch the indications on a blank drawing of the specimen. Despite additional efforts to clarify the coordinate system between the Level III and Level II testing, the coordinate system was a source of confusion throughout the round-robin inspections.

Table 31. Equipment used by Level II technicians during round-robin testing.

Technician Identifier	Instrument Type	Probe: No. of Elements	Frequency	Angles
1	Omniscan X3	16	5	45–70
2	Omniscan SX	32	5	40–70
3	Omniscan X3	32	5	40–72
4	Omniscan X3	16, 32	5	40–70
5	Omniscan X3	32	5, 10	40–70
6	Omniscan MX2	16	5	40–70
7	Omniscan SX	16	2.25, 5	40–72
8	Omniscan MX2	16	5	42–73
9	Omniscan MX2	16, 32	5	40–70
10	Omniscan MX3 / Epoch XT	16	5	38–72
11	Omniscan MX2	32	5	35–70
12	Omniscan X3	32, 128	5, 10	40–70
13	Omniscan MX2	16	5	36–72
14	Omniscan MX2	16, 32	5	37–72
15	Omniscan SX	16	5	40–70
16	Omniscan SX	16	7.5	40–70
17	Omniscan MX2	16, 64	5	40–70
18	Omniscan X3	16, 32	5	40–70

Similar to the Level III inspection trials, data from the Level II inspections were manually transferred from the inspection forms to a spreadsheet for evaluation. Results were compiled and summarized by technician instead of by specimen (Figure 24). Indications were associated with flaws using the same evaluation criteria as those used for the Level III inspection trials (Figure 17). Indications that were not associated with flaws were classified as false calls. The spreadsheet automatically summarized the results of the inspection (e.g., hits, misses, and false calls) and the error in flaw sizing (i.e., height and length) for each detected flaw. This summary was used in the data analysis.

Figure 24. Level II technicians during round-robin inspections.

4.3 Round-Robin Testing Results

Eighteen of the 21 technicians who were accepted based on their prequalification test results participated in one of the three five-day round-robin testing events held at CHL in Vicksburg, Mississippi, during February and March of 2022. Each technician scanned up to 21 welds, which included 68 flaws distributed across 20 welded specimens. One specimen incorporated two welds. The specimens included 12 ERDC specimens specifically designed by the research team to represent geometries and flaws encountered in HSS and 8 AWS D1.5 Bridge Kit specimens, which were designed for use in qualifying NDT technicians in the bridge industry. Results of the round-robin testing are enumerated by technician in Tables 32 and 33.

Table 32. Round-robin test results for ERDC specimens.

Technician Identifier	Flaws Scanned (*M*)	Flaws Detected			Total Missed (FN)	False Indications (FP)	Total Indications (PP)
		Flaws Sized		Total Detected (TP)			
		Within Tolerance (*S*)	Not within Tolerance (NS)				
1	44	16	9	25	19	8	33
2	44	20	10	30	14	7	37
3	44	26	7	33	11	5	38
4	44	23	12	35	9	21	56
5	44	30	3	33	11	11	44
6	44	30	5	35	9	5	40
7	44	33	2	35	9	2	37
8	44	28	7	35	9	4	39
9	44	30	5	35	9	3	38
10	44	32	4	36	8	4	40
11	44	29	5	34	10	5	39
12	41	32	4	36	5	5	41
13	44	36	1	37	7	2	39
14	44	36	1	37	7	5	42
15	44	31	5	36	8	3	39
16	44	33	4	37	7	1	38
17	44	37	1	38	6	3	41
18	44	33	7	40	4	2	42
Totals	789	535	92	627	162	96	723

ERDC and AWS specimens are reported separately in Tables 32 and 33, respectively. Table 34 reports the results for ERDC and AWS specimens

combined. The second column reports the total number of flaws (*N*) scanned by each technician. Some technicians scanned fewer flaws because, on average, they spent more time scanning each specimen. Technician 12 scanned fewer specimens because he was also asked to scan selected ERDC specimens using TOFD, but those results are not summarized in this report. The next three columns report the number of flaws that were scanned and detected, or TP results. TP results include flaws for which the technician's estimates of length and height were within the acceptable tolerances and those for which either length, height, or both were not within acceptable specified tolerances. Columns six through eight report the number of missed flaws, or false negatives (FN), the number of false indications, or FP results, and the number of total indications (PP).

Table 33. Round-robin test results for AWS D1.5 Bridge Kit specimens.

Technician Identifier	Flaws Scanned (*M*)	Flaws Detected and Sized			Total Missed (FN)	False Indications (FP)	Total Indications (PP)
		Flaws Sized		Total Detected (TP)			
		Within Tolerance (*S*)	Not within Tolerance (NS)				
1	24	19	2	21	3	1	22
2	21	16	2	18	3	3	21
3	24	20	0	20	4	4	24
4	24	17	3	20	4	16	36
5	18	16	2	18	0	7	25
6	24	20	1	21	3	2	23
7	24	21	0	21	3	2	23
8	15	13	1	14	1	9	23
9	24	21	1	22	2	1	23
10	24	21	0	21	3	2	23
11	24	19	3	22	2	2	24
12	6	4	0	4	2	20	24
13	24	22	0	22	2	5	27
14	21	20	0	20	1	3	23
15	24	24	0	24	0	0	24
16	24	22	1	23	1	0	23
17	24	23	0	23	1	1	24
18	24	19	1	20	4	2	22
Total	393	337	17	354	39	80	434

Table 34. Round-robin test results, ERDC and AWS specimens combined, using PAUT.

Technician Identifier	Flaws Scanned (*M*)	Flaws Detected and Sized			Total Missed (FN)	False Indications (FP)	Total Indications (PP)
		Flaws Sized		Total Detected (TP)			
		Within Tolerance (*S*)	Not within Tolerance (NS)				
1	68	35	11	46	22	9	55
2	65	36	12	48	17	10	58
3	68	46	7	53	15	9	62
4	68	40	15	55	13	37	92
5	62	46	5	51	11	18	69
6	68	50	6	56	12	7	63
7	68	54	2	56	12	4	60
8	59	41	8	49	10	13	62
9	68	51	6	57	11	4	61
10	68	53	4	57	11	6	63
11	68	48	8	56	12	7	63
12	47	36	4	40	7	25	65
13	68	58	1	59	9	7	66
14	65	56	1	57	8	8	65
15	68	55	5	60	8	3	63
16	68	55	5	60	8	1	61
17	68	60	1	61	7	4	65
18	68	52	8	60	8	4	64
Total	1,182	872	109	981	201	176	1,157

4.3.1 Diagnostic Test Evaluation

A diagnostic test is designed to detect the presence of a condition that may or may not be present and is usually hidden or hard to assess. Test results that accurately assess the presence or absence of a condition are classified as TP and true negative (TN), respectively. More often than not, diagnostic tests are associated with some degree of error. Test results that indicate the condition is present when in fact absent are FP test results and represent Type I errors. Those that indicate the condition is absent when in fact present are FN test results and represent Type II errors. In the NDT industry, TP test results are known as hits, FP results are known as false calls, and FN test results are known as misses. Some NDT studies attempt to quantify TN results by identifying likely flaw locations or by dividing the welds into damaged and undamaged lengths (Swensson et al. 1977; Spencer 1996a; Graybeal et al. 2001). However, other authors do not enumerate

TN results (Gruber and Light 2002; Shaw 2002). This study did not attempt to enumerate TNs because there did not appear to be an unambiguous approach to discretizing undamaged lengths.

The effectiveness of a diagnostic test can be analyzed using a confusion matrix, illustrated in Table 35. A confusion matrix enumerates test results as TP, FN, FP, and TN. The total number of cases in the population or number of flaws scanned is P_C = TP + FN. The total number of positive test results is P_T = TP + FP. Several metrics are commonly calculated to describe the efficacy of a diagnostic test from the confusion matrix. The sensitivity (SEN) of the test is the true positive rate (TPR), which is the proportion of flaws scanned that are detected, TPR = TP/P_C. The precision, or positive predictive value (PPV) of the test, is the proportion of positive results in which the condition being diagnosed is in fact present: PPV = TP/P_T. In NDT, PPV is known as the hit-to-call ratio. The accuracy (ACC) of the test is the fraction of test results that properly conclude that the condition is either present or absent: ACC = (TP + TN)/(P_C + N_C). The FDR is the fraction of positive test results obtained when the condition is absent: FDR = FP/P_T = 1 – PPV. The specificity (SPC) is the TN rate, or the proportion of negative test results in which the condition of interest is absent: SPC = TN/N_c, where N_c is as defined in Table 35. Because TN is not enumerated for this study, SPC could not be calculated directly.

Table 35. Confusion matrix example.

Condition		Diagnostic Test Result: Positive Test	Diagnostic Test Result: Negative Test	Row Total
Actual condition	Flaw present	TP (Hits)	FN (Misses)	$P_C = TP + FN$
	Flaw absent	FP (False calls)	TN	$N_C = FP + TN$
Column total		$P_T = TP + FP$	$N_T = FN + TN$	$P_T + N_T = P_C + N_C$

4.3.2 Technician Performance Metrics

Three performance metrics were calculated from round-robin test results for each technician (Table 36). These technician-level performance metrics describe three distinct aspects of skill with respect to NDT testing. The metrics include TPR, PPV, and the fraction of detected flaws sized within tolerances (FST). TPR reflects the ability of a technician to detect flaws within weld specimens. PPV describes the efficiency with which a technician detects flaws and reflects the ability to avoid making false calls. FST reflects the technician's ability to size flaws within tolerances. FST is

calculated as FST = S/TP, where S is the number of flaws sized within tolerances.

Table 36. Performance qualification and round-robin test performance metrics.

Technician Identifier	Performance Qualification			AWS D1.5 Bridge Kit			ERDC Specimens			ERDC and AWS Specimens		
	TPR	FST	PPV	TPR	FST	PPV	TPR	FST	PPV	TPR	FST	PPV
1	0.67	1.00	1.00	0.88	0.91	0.96	0.57	0.64	0.76	0.68	0.76	0.84
2	1.00	0.83	1.00	0.86	0.89	0.86	0.68	0.67	0.81	0.74	0.75	0.83
3	0.67	1.00	0.75	0.83	1.00	0.83	0.75	0.79	0.87	0.78	0.87	0.86
4	0.83	0.80	0.67	0.83	0.85	0.56	0.80	0.66	0.63	0.81	0.73	0.60
5	0.83	1.00	0.86	1.00	0.89	0.72	0.75	0.91	0.75	0.82	0.90	0.74
6	1.00	1.00	1.00	0.88	0.95	0.91	0.80	0.86	0.88	0.82	0.89	0.89
7	1.00	1.00	1.00	0.88	1.00	0.91	0.80	0.94	0.95	0.82	0.96	0.93
8	0.83	1.00	0.86	0.93	0.93	0.61	0.80	0.80	0.90	0.83	0.84	0.79
9	1.00	1.00	1.00	0.92	0.96	0.96	0.80	0.86	0.92	0.84	0.90	0.93
10	0.83	0.60	0.86	0.88	1.00	0.91	0.82	0.89	0.90	0.84	0.93	0.91
11	0.83	0.80	0.86	0.92	0.86	0.92	0.77	0.85	0.87	0.82	0.86	0.89
12	0.88	1.00	0.80	0.67	1.00	0.17	0.88	0.89	0.88	0.85	0.90	0.62
13	0.83	0.60	0.86	0.92	1.00	0.82	0.84	0.97	0.95	0.87	0.98	0.89
14	0.83	1.00	0.86	0.95	1.00	0.87	0.84	0.97	0.88	0.88	0.98	0.88
15	1.00	1.00	1.00	1.00	1.00	1.00	0.82	0.86	0.92	0.88	0.92	0.95
16	0.83	1.00	0.86	0.96	0.96	1.00	0.84	0.89	0.97	0.88	0.92	0.98
17	1.00	1.00	1.00	0.96	1.00	0.96	0.86	0.97	0.93	0.90	0.98	0.94
18	1.00	1.00	1.00	0.83	0.95	0.91	0.91	0.83	0.95	0.88	0.87	0.94

4.3.3 Correlations among Performance Metrics

Correlations between performance metrics were calculated to address two questions. The first was whether or not AWS specimens should be used as a substitute for ERDC specimens in prequalification. AWS specimens are much cheaper, easier to obtain, and easier to transport. However, several characteristics of AWS specimens may make them less challenging than the ERDC specimens. In AWS specimens, for example, joint geometries are less complicated, the plates are not as thick, the welds have been ground flush, and the flaws within the welds may be more evenly distributed. ERDC specimens are more representative of geometries encountered in HSS. To address this question, we compared performance on the AWS and ERDC specimens during round-robin testing. We did not use actual performance qualification results because most technicians scanned very few specimens during performance qualification, making any conclusions

from that comparison difficult. Round-robin results from technician 12 were not included in the correlation because this technician tested fewer AWS specimens during the round-robin testing event.

Table 37 summarizes the Pearson correlation coefficient, r, between performance on the AWS and ERDC specimens. The correlation between the TPR for AWS and the TPR for ERDC specimens was $r = 0.19529$. This correlation was low, and the p-value was not significant ($p = 0.4528$) compared to a critical p-value = 0.05. This criteria for statistical significance was used throughout this report. This result suggests that a technician's ability to detect flaws in AWS specimens is not an indicator of the technician's ability to detect flaws in ERDC specimens. However, results showed that when a flaw has been detected in an AWS specimen, the technician's ability to size that flaw within the tolerances adopted for this study may be an indicator of the ability to size flaws in ERDC specimens. The correlation between the FSTs for AWS and ERDC specimens was moderately high and statistically significant, $r = 0.70853$ ($p = 0.0015$). The correlation between PPV on AWS specimens and PPV on ERDC specimens was weaker, $r = 0.61433$ ($p = 0.0087$), indicating that a high PPV on AWS specimens may not be a good indicator of a technician's PPV on ERDC specimens. Therefore, it is preferable to avoid using AWS specimens in performance qualification when specimens that are representative of HSS geometries are available.

Table 37. Pearson correlation coefficients for performance metrics calculated from round-robin scans of AWS and ERDC specimens.

Performance Metric	Description	Pearson Correlation Coefficient (r)	p-value	Number of Observations (n)
TPR	True positive rate	0.1953	0.4528	17
FST	Probability of sizing the flaw within tolerances given it has been detected.	0.7085	0.0015	17
PPV	Positive predictive value	0.6143	0.0087	17

The second question was whether or not the three metrics of technician performance, TPR, FST, and PPV, reflected unique aspects of performance with respect to NDT. If these metrics were correlated, then the three metrics may have been redundant. This question was addressed in Table 38, which lists the Pearson correlation coefficient for each pair of metrics, the p-value for the correlation statistic, and the number of observations, n. Correlations in Table 38 indicate that there was a relatively high and

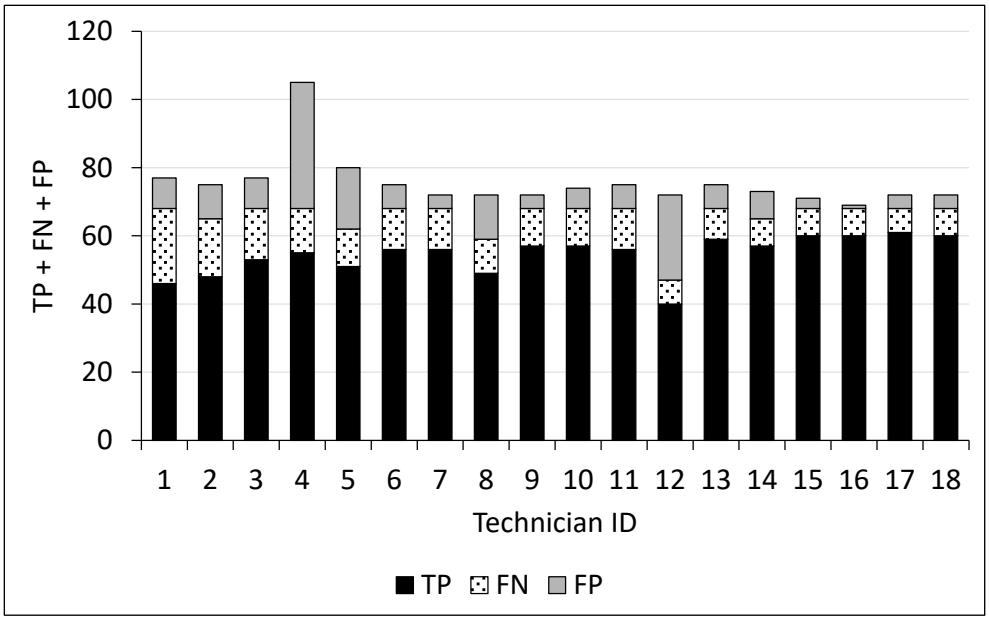
statistically significant correlation between TPR and FST for both AWS and ERDC specimens. Technicians who had higher TPRs were also more likely to estimate flaw length and height within tolerances. The correlation between TPR and PPV was positive and significant for ERDC specimens, indicating that technicians who had higher TPRs also tended to have higher PPVs. However, the correlation statistic was lower and not significant for AWS specimens. Similarly, the correlation between PPV and FST was positive, suggesting that technicians who had higher PPVs (i.e., lower rates of false discovery) were more likely to estimate flaw length and height within the specified tolerances. Again, for AWS specimens, the correlation was lower than it was for ERDC specimens and was not statistically significant for AWS specimens. The criteria for statistical significance used throughout this report was a p-value ≤0.05.

Table 38. Pearson correlation coefficients for performance metrics by specimen class.

Specimen Class	TPR versus FST			TPR versus PPV			FST versus PPV		
	r	*p*-value	*n*	*r*	*p*-value	*n*	*r*	*p*-value	*n*
AWS	0.78922	0.0002	17	0.1779	0.4945	17	0.4158	0.0969	17
ERDC	0.87131	<0.0001	17	0.5603	0.0193	17	0.6924	0.0021	17

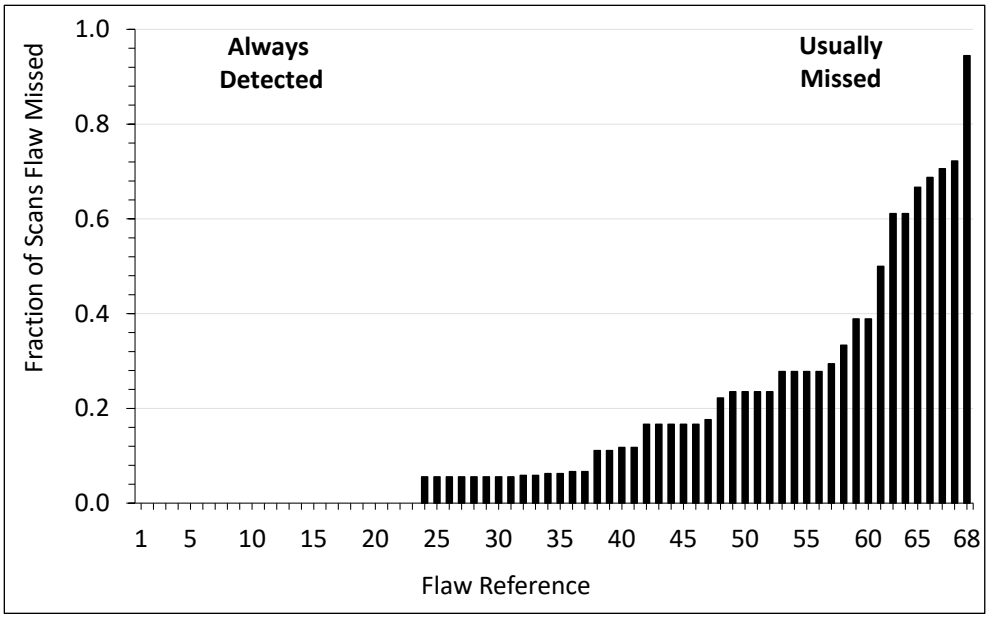
Round-robin test results showed that there was substantial variation in the skill exhibited by NDT technicians. Ideally, technicians would have high TPRs, FSTs, and PPVs. Figure 25 illustrates technician performance with respect to each of these metrics; it shows the number of hits (i.e., TP), misses (i.e., FN), and false calls (i.e., FP) for each technician. For each technician, test results were more likely to be classified as TP than as either FN or FP. Technicians 4 and 12 stood out as having a larger number of FPs than other technicians. Technician 4 scanned 68 flaws, missed 13 flaws, and reported 37 false indications. Technician 12 scanned 47 flaws using PAUT, missed 7 flaws, and reported 25 false indications. In contrast, technicians 15 and 16 were the best performers. Technician 16 scanned 68 flaws, missed 8 flaws, and reported only 1 false indication. Technician 15 scanned 68 flaws, missed 8 flaws, and reported 3 false indications.

Figure 25. Number of hits (*black*), misses (*stippled*), and false calls (*gray*) by technician for PAUT scans (ERDC and American Welding Society [AWS] specimens).



While technician skill is usually considered to be an important factor influencing the ability to detect and size flaws, several other factors may also influence a technician's ability to detect and size flaws. These factors include the characteristics of the flaw (e.g., dimensions, flaw category, and orientation) and the characteristics of the joints and plates in which the flaws are situated (e.g., joint type, the presence of skewed members in the joint, plate thickness, and weld reinforcement). The location of a flaw within a specimen and in relation to other flaws within that specimen may also be important. This is illustrated by the round-robin test results, which showed that while many flaws were consistently detected, some flaws were usually missed regardless of technician. These results are shown in Figure 26, which shows a bar graph that describes the frequency with which each of the 68 round-robin flaws was detected. Each flaw is identified by its flaw reference code on the x-axis. The fraction of times each flaw was missed by a technician who scanned it is shown on the y-axis. For example, flaw 68 was detected in only 1 of the 18 scans (5.56%), and flaw 67 was detected in only 5 of 18 scans (27.8%). In contrast, many flaws were always detected. For example, 23 of the 68 flaws were detected in all of the scans.

Figure 26. Fraction of PAUT scans during which the flaws were missed.



4.3.4 Logistic Regression Models for Probability of Detection (POD)

Statistical models are useful tools for quantifying the effects of multiple flaw and specimen characteristics on a technician's ability to detect flaws because they can account for the effects of multiple factors simultaneously. In this study, logistic regression models were used to estimate the effects of flaw and specimen characteristics on POD. The dependent variable, y, in a logistic regression is a binary variable that takes the value 1 if an objective criteria is met and takes a value of 0 (zero) otherwise. A logistic regression equation estimates the probability that a binary dependent variable, $y \in [0,1]$ is equal to 1 given a vector of independent variables, x_i:

$$p(y = 1|\boldsymbol{x}) = \frac{\exp(\alpha+\sum_{i=1}^{N}\beta_i x_i)}{1+\exp(\alpha+\sum_{i=1}^{N}\beta_i x_i)}.$$

The $i \in \{1,2, \ldots, N\}$ independent variables on the right-hand side of the equation describe the flaw and specimen characteristics of interest, which may be continuous, discrete, or categorical in nature. The intercept, a, and parameters, β_i, are estimated from the logistic regression. In the result sections that follow, two models were estimated for each dependent variable. The first was the initial model run that included all of the independent variables initially proposed for analysis. The second included only those independent variables with parameter estimates that satisfied the significance criteria, $p \leq 0.05$.

Table 39 describes the independent variables considered in this study. Flaw characteristics of interest included length, height, and aspect; for laminar flaws, however, area was used instead of height. Flaw length is the distance between the start and stop location of each flaw along the *y*-dimension. Flaw height is the distance between the start and stop locations of each flaw in the *z*-dimension. Flaw aspect is the ratio of flaw height to flaw length. Although flaw orientation is often an important factor that influences detection, it was not considered in this study because technicians were instructed to manually raster scan the specimens from all sides. This limited the effect of orientation on detectability and made flaw orientation difficult to define.

Table 39. Independent variables used in logistic regression by flaw category.

Variable	Description	Units	Domain	Range or Levels
Length	Flaw length along the *y*-axis	in.	Continuous	See Table 41
Height	Flaw length along the *z*-axis	in.	Continuous	See Table 41
Aspect	Ratio of height to length	—	Continuous	See Table 41
Area	Product of flaw length and width	$in.^2$	Continuous	See Table 41
Thickness	Material thickness of the thicker part joined	in.	Continuous	See Table 41
TPR	True positive rate	—	Continuous	See Table 41
Precision	Positive predictive value	—	Continuous	See Table41
Instrument	Olympus Omniscan Model	—	Categorical	X3, SX, MX2, MX3
Joint type	Joint type	—	Categorical	B: Butt, C: Corner, T: T-joint
Joint skew	Presence of a skewed joint member	—	Binary	0: False, 1: True
Flaw category	Flaw category	—	Categorical	P: Planar, V: Volumetric, L: Laminar
Flaw subcategory	Flaw subcategory	—	Categorical	See Table 40
Class	Weld specimen origin	—	Categorical	AWS: Commercial ERDC: Custom

Flaws were also distinguished by category (i.e., planar, volumetric, or laminar) and subcategory (Table 40). Specimen characteristics considered in this analysis included joint type (i.e., butt, corner, or T) and the presence of skewed members in the joint (i.e., joint skew). Corner joints, T-joints, and in particular skewed joints limited physical access for the probe, making it more difficult to distinguish between reflections caused by flaws and those caused by geometry and more difficult to determine flaw location. These factors may have contributed to errors in flaw

detection and sizing. The instrument type and the overall level of skill demonstrated by the technician relative to other round-robin participants were also of interest. Depending upon the dependent variable, this was represented using one of two round-robin performance metrics calculated for each technician.

Table 40. Levels for flaw category and subcategory.

Flaw Category	Flaw Subcategory (Abbreviation: Description)
Planar (P)	ROCRK: root crack, TOCRK: toe crack, CLCRK: centerline crack, TRCRK: transverse crack, BMCRK: base metal crack, LOF: lack of fusion, LOP: lack of penetration
Volumetric (V)	POR: porosity, SLAG: slag
Laminar (L)	LAM: lamination

For detection, the technician's TPR was used as a relative measure of skill. For indication reliability, the technician's PPV was used as a relative measure of skill. Table 41 summarizes the descriptive statistics for all model covariates. The minimum value of 0.01 in. that is reported for height in Table 41 is for a planar flaw for which the height was undefined in as-built drawings.

Factors that could potentially influence the probability of flaw detection were evaluated by fitting the logistic regression model to a dichotomous dependent variable that was equal to 1 if the technician detected the flaw and was equal to 0 otherwise. Logistic regressions were fit to three independent subsets of the round-robin test results. Round-robin data were segregated based on flaw category because, in initial exploratory models, this led to substantial improvements in model performance overall. In addition, not all independent variables were applicable to all flaw categories. For example, height and aspect were not defined for laminar flaws. The size of laminar flaws was characterized as an area. Tables 42 to 44 summarize the results for each flaw category. The effect coding method was used for categorical variables to compare each level of that variable to a reference level (SAS Institute, n.d.). One level of each categorical variable was assigned as the reference level, and one independent variable was created for each of the remaining levels. POD estimates were obtained for each nonreference level by assigning the value 1 to the variable for that level and the value 0 to each of the remaining levels for that categorical variable. When estimating POD for the reference level, a value of –1 was assigned to each of the variables for the nonreference levels.

Table 41. Descriptive statistics for continuous independent variables used in logistic regression by flaw category.

	Flaw Category (*n* = Number of Observations)											
	Planar (*n* = 836)				Volumetric (*n* = 275)				Laminar (*n* = 71)			
Independent Variables	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Length (in.)	1.0547	0.9259	0.0800	4.5200	1.4183	1.6256	0.2800	6.9800	0.8773	0.3902	0.3500	1.4100
Height (in.)	0.1878	0.1177	0.0100	0.5400	0.1433	0.0594	0.0500	0.2600	—	—	—	—
Aspect	0.2670	0.2036	0.0000	1.0000	0.2054	0.1729	0.0158	0.7122	—	—	—	—
Area (in.^{2})	—	—	—	—	—	—	—	—	0.9285	0.7109	0.1190	2.0000
Plate thickness (in.)	1.0952	0.8337	0.5000	4.0000	1.0317	0.8278	0.5000	4.0000	0.8256	0.4048	0.5100	1.5100
TPR	0.8299	0.0544	0.6765	0.8971	0.8300	0.0547	0.6765	0.8971	0.8299	0.0545	0.6765	0.8971
PPV	0.8600	0.1027	0.5978	0.9836	0.8630	0.1004	0.5978	0.9836	0.8586	0.1033	0.5978	0.9836

4.3.4.1 Planar Flaws

Table 42 summarizes the logistic regression results for the detectability of planar flaws. Planar flaw subcategories include LOF, LOP, and five different cracks: BMCRKs, CLCRKs, ROCRKs, TOCRKs, and TRCRKs. The table lists the independent variables and parameter estimates; for categorical variables, it also lists the respective levels associated with each parameter estimate. Parameter estimates for categorical variables are given relative to a reference level, which is noted as (Ref =) for each variable. For example, the variable flaw subcategory has seven levels. Parameter estimates are listed for six levels, and the reference category is TRCRK. There is no parameter estimate for the reference category. Negative parameter estimates indicate that detectability decreases as the covariate increases or given the presence of a condition. Positive parameter estimates indicate a positive effect on detectability.

The Wald chi-square statistic and the *p*-value for that statistic (pr > chi square) describe the statistical significance of the parameter estimates. By convention, parameter estimates are statistically significant if the *p*-value ≤ 0.05. Three independent variables were not significant: flaw length, joint type, and specimen class. The insignificance of the parameter estimate on flaw length was surprising given that many POD studies emphasize flaw length as a factor influencing POD, and it is generally assumed that larger flaws are easier to detect. While the flaw length variable was dropped from the revised model, flaw length was still accounted for in the model by the flaw aspect variable. Flaw aspect is the ratio of flaw height to flaw length and, while flaw aspect is a physical property of flaws, this variable can be considered an interaction term. As with flaw length, joint type was also expected to be a significant factor influencing POD because it is easier to access a butt joint from all sides than it is to access T- and corner joints from all sides. While the parameter estimates were not significant in the initial model, they did have the expected signs. The reference category was the T-joint (T). The parameter estimate was positive for butt joints and negative for corner joints. The *p*-value for the butt joint parameter estimate, 0.1443, was almost significant, and the *p*-value criterion, although widely used and accepted, is somewhat arbitrary.

Table 42. Logistic regression results for probability of detection (POD) of planar flaws.

Model Version		Initial Model				Revised Model			
Parameter	Discrete Level	Param. Est.	Std. Error	Wald Chi-Square	Pr > Chi-Square	Param. Estimate	Std. Error	Wald Chi-Square	Pr > Chi-Square
Intercept	—	−4.8546	1.7276	7.8961	0.005	−4.8602	1.5348	10.0278	0.0015
Flaw length (in.)	—	0.0877	0.2446	0.1286	0.7199	—	—	—	—
Flaw height (in.)	—	3.7698	1.5803	5.6906	0.0171	3.0382	1.2081	6.3248	0.0119
Flaw aspect	—	−1.1142	0.6644	2.8121	0.0936	−1.3516	0.4980	7.3668	0.0066
Plate thickness (in.)	—	−0.8578	0.1477	33.7113	<.0001	−0.8899	0.1264	49.5587	<.0001
Technician TPR	—	7.799	2.027	14.8035	0.0001	8.1375	1.8290	19.7954	<.0001
Omniscan instrument (Ref = X3)	MX2	0.0738	0.1981	0.1387	0.7096	—	—	—	—
	MX3	−0.0016	0.354	0	0.9964	—	—	—	—
	SX	−0.0136	0.2141	0.004	0.9493	—	—	—	—
Joint type (Ref = T)	B	0.2831	0.1933	2.1444	0.1431	—	—	—	—
	C	−0.0966	0.2406	0.1613	0.688	—	—	—	—
Joint skew (Ref = TRUE)	FALSE	0.7287	0.257	8.0385	0.0046	0.8812	0.2249	15.3504	<.0001
Flaw category (Ref = TRCRK)	BMCRK	−0.9919	0.3402	8.5	0.0036	−1.2203	0.3008	16.4613	<.0001
	CLCRK	0.3331	0.3117	1.1423	0.2852	0.3159	0.3040	1.0802	0.2986
	LOF	0.4611	0.2965	2.4177	0.12	0.4442	0.2849	2.4301	0.119
	LOP	1.5611	0.4553	11.7585	0.0006	1.7943	0.4468	16.1278	<.0001
	ROCRK	−1.2912	0.2794	21.3612	<.0001	−1.3425	0.2681	25.0644	<.0001
	TOCRK	1.1357	0.371	9.3728	0.0022	1.2237	0.3266	14.0354	0.0002
Class (Ref = ERDC)	AWS	0.1195	0.1701	0.4938	0.4822	—	—	—	—
Model fit statistics		Likelihood ratio chi-square = 168.2882 (p < 0.00001, df = 18)				Likelihood ratio chi-square = 165.3535 (p < 0.00001, df = 11)			
		Akaike information criterion (AIC), intercept: 731.263 AIC, intercept, and covariates: 598.975				AIC, intercept: 731.263 AIC, intercept, and covariates: 587.910			
		Percent concordant: 82.3% Area under the receiver operator curve (ROC) curve: 0.8227				Percent concordant: 82.1% Area under the ROC curve: 0.8212			
Response variable		y = 1: flaw detected; 0 = flaw not detected							
Number of observations		n = 836 ($n_{y=1}$: 704; $n_{y=0}$: 132)							

A revised logistic regression model was fit to the planar flaw dataset, excluding the insignificant variables. Dropping the insignificant variables increased the significance of the remaining variables and, in the case of flaw aspect, reduced the p-value below its critical level. In the revised model, technician skill, represented by the individual's TPR, had the single biggest effect on POD. Increasing flaw height also increased detectability. However, the negative coefficient on flaw aspect indicated that, as height increased, there was a diminishing effect on POD. All else equal (i.e., as long

as all of the other variables remain the same), increasing the length of a flaw with a given height will reduce the aspect and increase the POD. Increasing plate thickness and joint skew both reduce POD. Parameter estimates for BMCRK and ROCRK were negative, suggesting that these flaw categories were more difficult to detect than TRCRK, which was the reference level. Conversely, the parameter estimates for LOF, LOP, and TOCRK were all positive and statistically significant, suggesting that these categories were easier to detect than TRCRK. The parameter estimate for CLCRK was positive but not significant, indicating that, statistically, it was no more or less detectable than TRCRK. Insignificant parameter estimates for one or more levels of categorical variables were retained in the model if the parameter estimate for at least one other level was significant. Instrument type was not significant.

The overall validity of the logistic regression model for planar flaws is supported by the stability of the parameter estimates across exploratory changes to the model and model fit statistics. The stability of the parameter estimates is demonstrated by comparing the estimates in the initial and revised model. While some sensitivity was expected, the removal of insignificant variables did not result in dramatic changes in those parameter estimates. Model fit statistics and the number of observations in the planar flaw dataset are summarized at the bottom of Table 42. The likelihood ratio chi-square tests were significant, and the Akaike information criterion (AIC) was slightly lower in the revised model than in the initial model. The revised model was more parsimonious than the initial model. Percent concordance describes the fraction of pairs of observations for which the estimates of POD are aligned with the observed detection frequencies. The area under the receiver operator characteristic (ROC) curve describes the probability that the model accurately predicts whether a flaw will be detected or missed. A model with a value of 1 predicts perfectly, and a model with a value of 0.5 predicts randomly.

4.3.4.2 Volumetric Flaws

Table 43 summarizes the logistic regression results for the detectability of volumetric flaws. Volumetric flaws include two flaw categories, SLAG and POR. Parameter estimates for flaw length, aspect, flaw category, and specimen class were significant in the initial model. The coefficient on flaw length was negative, suggesting that POD decreases as flaw length increases. Flaw aspect had the largest influence on POD. Parameter estimates for flaw height and technician TPR were marginally significant in

the initial model and had a positive effect on POD. Although it was not significant in the initial model, it appeared to be significant in the revised model. The coefficient on plate thickness was positive, indicating that, all else equal, flaws in thicker plates are easier to detect. Although technician TPR did not satisfy the significance criteria of $p \leq 0.05$ for inclusion in the final model, this variable was retained because of its perceived importance as an explanatory variable. Instrument type was not significant.

Table 43. Logistic regression results for POD of volumetric flaws.

Model Version		Initial Model				Revised Model			
Parameter	Discrete Level	Param. Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Param. Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
Intercept	—	−9.6484	41.7081	0.0535	0.8171	−5.1302	2.989	2.9458	0.0861
Flaw length (in.)	—	−0.3657	0.1315	7.7339	0.0054	−0.3178	0.124	6.5688	0.0104
Flaw height (in.)	—	8.4942	5.5816	2.3159	0.1281	—	—	—	—
Flaw aspect	—	6.9633	2.5516	7.4473	0.0064	7.7005	2.3981	10.3109	0.0013
Plate thickness (in.)	—	0.9697	0.5424	3.1959	0.0738	1.1232	0.4275	6.9028	0.0086
Technician TPR	—	7.4818	4.0365	3.4356	0.0638	5.4417	3.5035	2.4125	0.1204
Omniscan instrument (Ref = X3)	MX2	−0.3485	0.3585	0.9446	0.3311	—	—	—	—
	MX3	−0.2537	0.6419	0.1562	0.6927	—	—	—	—
	SX	0.3847	0.4169	0.8516	0.3561	—	—	—	—
Joint type (Ref = T)	B	−2.0398	83.1227	0.0006	0.9804	1.5972	0.5986	7.1207	0.0076
	C	−4.643	83.1217	0.0031	0.9555	−0.8546	0.4064	4.4208	0.0355
Joint skew (Ref = TRUE)	FALSE	5.69	124.7	0.0021	0.9636	—	—	—	—
Flaw category (Ref = SLAG)	POR	−1.563	0.3337	21.9349	<.0001	−1.4252	0.2969	23.0372	<.0001
Class (Ref = ERDC)	AWS	1.3614	0.4008	11.5362	0.0007	1.4892	0.3945	14.2472	0.0002
Model fit statistics		Likelihood ratio chi-square = 100.7937 (p < 0.00001, df = 13)				Likelihood ratio chi-square = 94.5816 (p < 0.00001, df = 9)			
		AIC, intercept: 259.744 AIC, intercept, and covariates: 184.950				AIC, intercept: 259.744 AIC, intercept, and covariates: 181.162			
		Percent concordant: 88.9% Area under the ROC curve: 0.8896				Percent concordant: 87.5% Area under the ROC curve: 0.8762			
Response variable		1 = flaw detected; 0 = flaw not detected							
Number of observations		n = 275 ($n_{y=1}$: 226; $n_{y=0}$: 49)							

There were at least three potential concerns surrounding the logistic regression results for volumetric flaws. The negative coefficient on length was not consistent with the usual hypothesis that longer flaws are easier to detect. As some of the Level III technicians explained during the validation of NDT procedures, indications associated with longer flaws were often attributed to signal created by geometry because the start and stop locations

were not obvious and because technicians were not expecting to see flaws of such great length in NDT specimens. However, this coefficient remained negative even after the three longest flaws were removed from the dataset. The second potential concern was that the sign on the plate thickness coefficient was opposite the sign for the plate thickness coefficient in the logistic regression for planar flaws. This may reflect the complexity of using plate thickness as an explanatory variable. Flaws in thinner plates (e.g., 5/16 in.) are generally considered to be harder to detect than flaws in thicker plates, but this is true only to a point (e.g., 1 in.). Flaws in plates that are thicker than 1 in., like the 4 in. plates used in this study, may be harder to detect because the ultrasonic sound path is longer. Finally, the coefficient on technician TPR was not significant. These three issues may reflect the sparseness of data used to fit the logistic regression to a large number of independent variables with multiple levels. Only 275 observations were available to fit the logistic regression for volumetric flaws, which is just one-third of the 832 observations available to fit the logistic regression for planar flaws.

4.3.4.3 Laminar Flaws

Table 44 summarizes the logistic regression results for the detectability of laminar flaws. The independent variables included lamination area, plate thickness, technician TPR, and joint type. There were no laminations in corner joints, so only butt joints and T-joints are represented in the model. Parameter estimates for lamination area and plate thickness were not significant in the initial model. However, the parameter estimate for lamination area became significant after plate thickness was dropped from the model. The parameter estimate for technician TPR was significant, indicating that technician skill is an important factor influencing the detectability of laminations. Instrument type was not significant. Specimen class was not included in this logistic regression model because AWS specimens contained no laminations. Only 71 observations were available to fit the logistic regression models.

Table 44. Logistic regression results for POD of laminar flaws.

Model Version		Initial Model				Revised Model			
Parameter	Discrete Level	Param. Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Param. Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
Intercept	—	−22.6478	7.0148	10.4237	0.0012	−21.5606	6.4079	11.3210	0.0008
Lamination area (in.^{2})	—	0.1065	1.9359	0.003	0.9561	1.9584	0.5927	10.9186	0.0010
Plate thickness (in.)	—	3.9055	3.8275	1.0412	0.3075	—	—	—	—
Technician TPR	—	24.9716	7.9345	9.9049	0.0016	25.5427	7.5878	11.3319	0.0008
Omniscan instrument (Ref = X3)	MX2	0.4637	0.6994	0.4396	0.5073	—	—	—	—
	MX3	−0.0157	1.2181	0.0002	0.9897	—	—	—	—
	SX	−0.6236	0.7084	0.775	0.3787	—	—	—	—
Joint type (Ref= T)	B	1.4114	0.586	5.8007	0.016	0.9894	0.3791	6.8107	0.0091
Model fit statistics		Likelihood ratio chi-square = 36.8411 (p < 0.00001, df = 7)				Likelihood ratio chi-square = 34.4889 (p < 0.00001, df = 3)			
		AIC, intercept: 86.425 AIC, intercept, and covariates: 63.584				AIC, intercept: 88.425 AIC, intercept, and covariates: 57.936			
		Percent concordant: 91.2% Area under the ROC curve: 0.9137				Percent concordant: 89.8% Area under the ROC curve: 0.9005			
Response variable		1 = flaw detected; 0 = flaw not detected							
Number of observations		n = 71 ($n_{y=1}$: 51, $n_{y=0}$: 20)							

Several critical assumptions should be satisfied to ensure that logistic regression parameter estimates are unbiased and stable because bias and instability can lead to large confidence bounds and invalid statistical inferences. Key assumptions include (1) there are no correlations among the independent variables (multicollinearity), (2) the relationship between continuous independent variables and the log odds ratio is linear, (3) there are no strongly influential outliers, and (4) the sample size is sufficiently large. These assumptions are always satisfied to a degree. The absence of a correlation between independent variables was confirmed, and diagnostic information was reviewed, to investigate the linearity between independent variables and the log odds ratio and to identify strongly influential outliers. Particularly with respect to fitting logistic regression models to the volumetric and laminar flaw datasets, a larger number of observations could help improve the models. One general rule says that at least 10 observations of the least-frequent outcome are needed for each independent variable in a logistic regression model. For example, the revised logistic regression for planar flaws contained 11 independent variables, and the least-frequent outcome (i.e., miss) occurred with a relative frequency of

0.158. None of the initial models satisfy this rule. With respect to the revised models, the minimum number of observations needed for logistic regression was approximately $11 \cdot 10/0.158 = 696$. This was very close to the 837 observations in the planar flaw dataset. Based on this rule, logistic regression on the volumetric flaw dataset required 450 observations (there are 275), and logistic regression on the laminar flaw dataset required 101 observations (there are 71). This suggests that all of the models might be improved by having a larger number of observations to fit them. However, that is almost always the case.

4.4 POD Model Sensitivity

Logistic regression models estimate the POD while accounting for selected flaw and specimen characteristics and technician skill. The objective of sensitivity analysis is to illustrate the influence of flaw and specimen characteristics by varying them one at a time. Two caveats are necessary. The first is that, while comparisons within a given flaw category are valid, comparisons across different flaw categories are not. Each flaw category is represented by a distinct logistic regression model that has been fit to an independent dataset. The second caveat is that the datasets are relatively sparse, and there are a large number of categorical variables and levels. Some combinations of flaw and specimen characteristics are not fully represented in the round-robin datasets, and these gaps amount to blind spots. Conclusions based on these models accurately reflect the observations made during these round-robin testing experiments. However, readers should exercise caution when extrapolating these conclusions to other settings.

In Table 45, POD estimates are summarized for planar, volumetric, and laminar flaw categories. Near the top of the table, a nominal POD estimate is included; it was calculated with all continuous independent variables at their mean value and all categorical variables assigned a value of 0 (zero). This represents a nominal estimate of POD for each flaw category. This nominal flaw is a device for describing the central tendency of the data. It is a value that might be used as an estimate of the POD when information on flaw and specimen characteristics is not available. In subsequent rows of the table, the POD was calculated varying each flaw and specimen characteristic one at a time. When estimating a POD for the reference level of a categorical variable, a value of –1 is assigned to each level of that variable, while a value of 0 (zero) is assigned to all levels of other categorical

variables. When estimating a POD for a nonreference level of a categorical variable, a value of 1 is assigned to that level of the categorical variable, and a value of 0 (zero) is assigned to all other levels of that variable and all other categorical variables.

Table 45. Sensitivity of detection probability to significant independent variables.

Factor	Case Description	POD by Flaw Category		
		Planar	Volumetric	Laminar
None	Nominal case	0.7548	0.8424	0.8109
Technician TPR	Lowest TPR	0.4691	0.6986	0.0785
	Highest TPR	0.8418	0.8851	0.9598
Joint type	Butt joint (B)	—	0.9635	0.9202
	Corner joint (C)	—	0.6946	—
	T-joint (T)	—	0.7178	0.6145
Joint skew	FALSE	0.8814	—	—
	TRUE	0.5607	—	—
Flaw category or subcategory	Transverse crack (TRCRK)	0.4773	—	—
	Base metal crack (BMCRK)	0.4761	—	—
	Centerline crack (CLCRK)	0.8085	—	—
	Root crack (ROCRK)	0.4457	—	—
	Toe crack (TOCRK)	0.9128	—	—
	Lack of fusion (LOF)	0.8276	—	—
	Lack of penetration (LOP)	0.9488	—	—
	Porosity (POR)	—	0.5624	—
	Slag (SLAG)	—	0.9569	—
Specimen class	AWS	—	0.9595	—
	ERDC	—	0.5466	—

As an example, the nominal POD for planar flaws is described here. Eighteen NDT technicians each conducted PAUT scans of up to 48 planar flaws during round-robin testing. This generated 836 observations, which was less than the 864 observations that would be expected if all technicians had scanned all specimens. Averaging over all planar flaw scans, the average planar flaw had a length of 1.055 in., a height of 0.188 in., and an aspect of 0.2670. The average flaw was situated in a specimen with a maximum plate thickness of 1.1 in. The average technician TPR was 0.8299. The POD for the nominal flaw was estimated assuming these mean values. For the nominal case, the categorical variables were left undefined (evaluated at 0). Table 45 lists the POD for the nominal flaw of

each flaw category, and in the first column, “none” means that no factors are being varied in the nominal case.

Technician skill can have a large effect on POD (Table 45). The effect of technician skill was assessed by varying technician TPR from 0.6765, representing the technician with the lowest TPR, to 0.8971, representing the technician with the highest TPR (Table 41). For technician 1, with the lowest TPR, the POD decreased to 0.4691. For technician 17, with the highest TPR, the POD increased to 0.8418. In other words, the technician with the highest TPR was 80% more likely to detect the nominal flaw than the technician with the lowest TPR and 12.8% more likely to detect the nominal flaw than the technician with the average TPR.

The sensitivity of POD to flaw and specimen characteristics is assessed by varying the categorical independent variables of the logistic regression one at a time. For example, joint skew has two levels, and the reference level is true. If joint skew is absent (i.e., joint skew = false), the POD is estimated by assigning a value of 1 to the independent variable for joint skew. If joint skew is present (i.e., joint skew = true), the POD is estimated by assigning a value of –1 to the independent variable for joint skew. All other categorical variables are assigned a value of 0. The sensitivity of the nominal flaw POD to joint skew is the difference between the POD without joint skew and the POD with joint skew.

For planar flaws, the significant categorical variables were joint skew and flaw subcategory. The POD for planar flaws without joint skew was 0.8814. The POD for planar flaws with joint skew was 0.5607. The nominal effect of joint skew was to reduce the POD by 36.3%. Of course, the size of this effect varied across combinations of flaw and specimen characteristics. Flaw subcategories were also significant factors influencing the POD of planar flaws. Planar flaw subcategories included LOF, LOP, and five different types of cracks (i.e., TRCRK, BMCRK, CLCRK, ROCRK, and TOCRK). For TRCRK, the reference level of the subcategory variable, POD was 0.4773. This was 37% less than the POD of the nominal flaw. Planar flaws in the BMCRK and ROCRK subcategories were also less likely to be detected than the nominal flaw. Planar flaws in other subcategories were more likely to be detected than the nominal flaw.

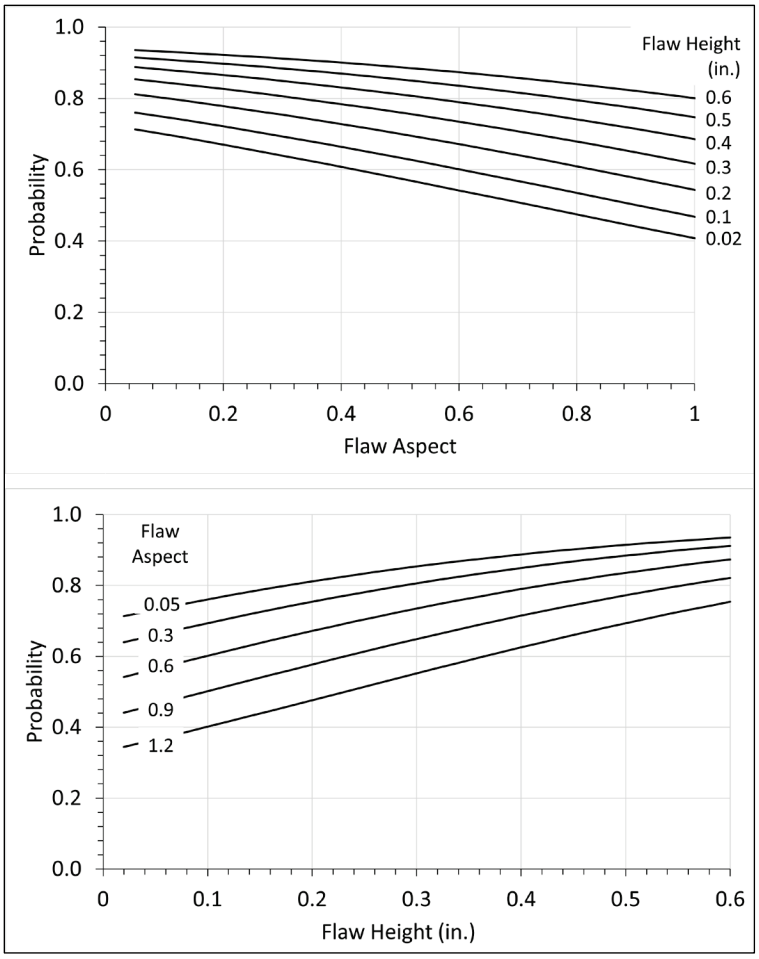
For volumetric flaws, the nominal flaw had a POD of 0.8424. Significant categorical variables included joint type, flaw subcategory, and specimen

class. Volumetric flaws in butt joints (POD = 0.9635) were easier to detect than those in T- and corner joints (POD = 0.6946 and POD = 0.7178, respectively). Joint skew was not a significant factor in the logistic regression for volumetric flaws. Flaws characterized as SLAG (POD = 0.9569) were easier to detect than those characterized as POR (POD = 0.5624). Volumetric flaws in ERDC specimens (POD = 0.5466) were more difficult to detect than volumetric flaws in AWS specimens (POD = 0.9595).

For laminar flaws, fewer observations were available to fit logistic regressions because there were only four laminar flaws in the round-robin experiment. The nominal laminar flaw had a POD of 0.8109. The single greatest factor determining the POD for laminar flaws was technician TPR. Joint type was also a significant categorical variable. Laminar flaws were more difficult to detect in T-joints (POD = 0.6145) than in butt joints (POD = 0.9202). There were no laminations in corner joints or in AWS specimens.

Figures 27 through 29 illustrate the sensitivity of POD for selected continuous independent variables with the largest coefficients. For planar flaws, the POD was influenced by flaw height (Figure 27). Increases in flaw aspect while holding flaw height constant reduced the POD for planar flaws. This was equivalent to reducing flaw length and was the expected response. The other significant continuous variables were plate thickness and technician TPR. The coefficient on plate thickness was negative, suggesting that planar flaws in thicker plates were more difficult to detect. While, all else equal, flaws in thinner plates (e.g., 5/16 in.) are usually considered harder to detect than flaws in thicker plates up to about 1 in., flaws in very thick plates (e.g., 4 in.), like some of those used in this study, may be harder to detect because the ultrasonic sound path is longer. Increasing technician TPR had a positive effect on POD. These effects are not shown in the Figure 27.

Figure 27. Probability of detection (POD) for the nominal planar flaw when varying flaw height and flaw aspect.



For volumetric flaws, the significant continuous independent variables were flaw length, aspect, and plate thickness. POD increased with flaw aspect and plate thickness. These results show that, all else equal, flaws with a greater ratio of height to length were easier to detect than flaws with lower ratios of height to length and that flaws in thicker plates were easier to detect than those in thinner plates (Figure 28). However, for those plates that were much thicker than about 1 in., it was expected that plate thickness would reduce the POD because, as the length of the sound path increases, the attenuation of ultrasonic energy increases. This expectation was not reflected in these results. When examining these results, it is important to remember that flaw aspect depends on both length and height. If a 2 in. flaw has an aspect of 0.1, then it must have a height of 0.2 in. All else equal, volumetric flaws with aspects greater than 0.4 and lengths greater than 4 in. have PODs approaching 1. This reflects the data collected during the round-robin experiments and under ideal testing conditions

(e.g., the technicians were working indoors in a climate-controlled environment and with good lighting).

Figure 28. POD for the nominal volumetric flaw when varying plate thickness and flaw aspect.

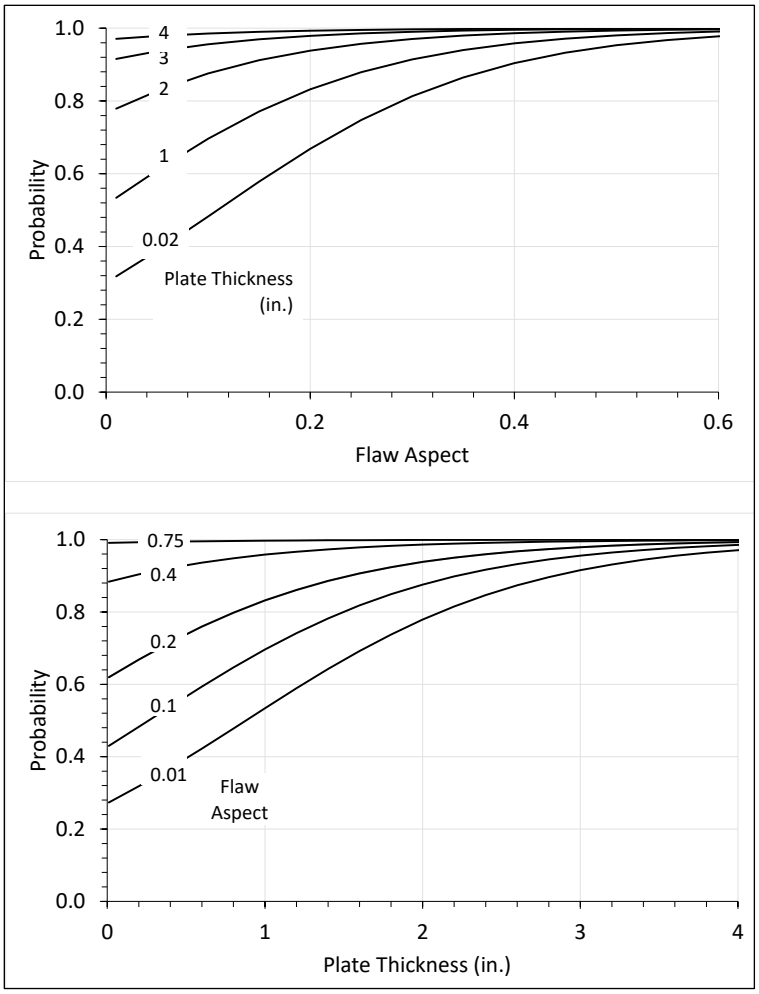
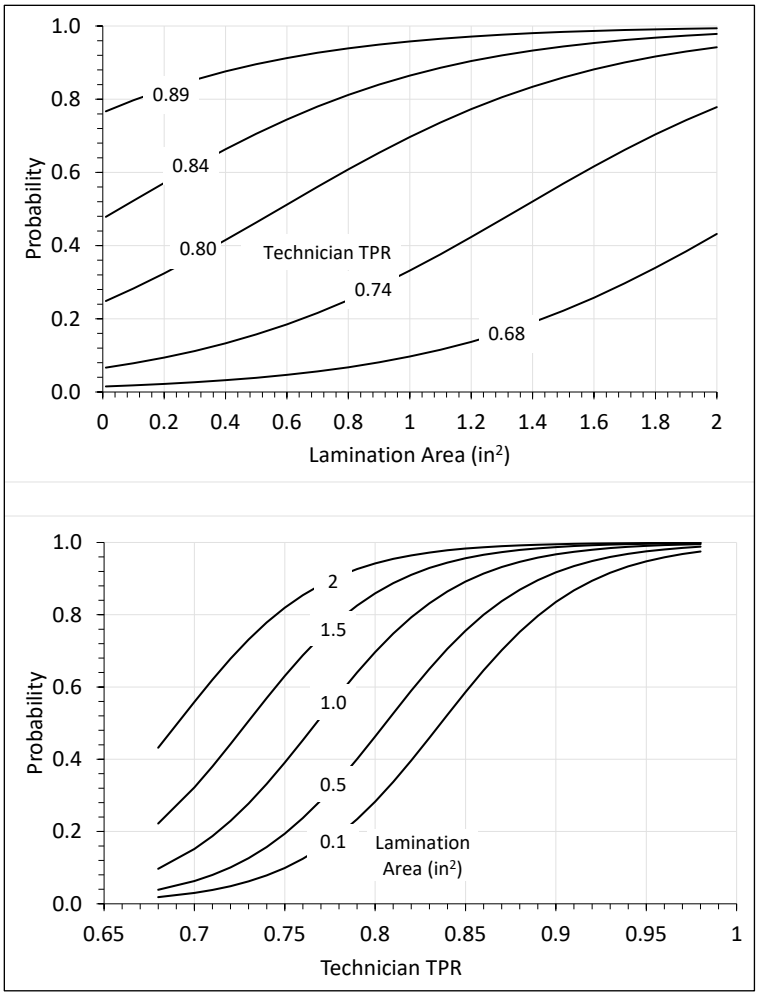


Figure 29 summarizes the results for laminar flaws. The significant continuous independent variables were lamination area and technician TPR. Lamination area is the product of the lamination length, measured along the *y*-axis, and the lamination width, measured along the *x*-axis. Laminations are two-dimensional flaws with an undefined height. Both lamination area and technician TPR exhibited a strong positive relationship with POD. The POD increased with both increasing lamination area and increasing technician TPR. Technicians with the lowest TPRs were more likely to miss a lamination than to detect it.

Figure 29. POD for the nominal laminar flaw when varying technician true positive rate (TPR) and lamination area.



4.5 POD Estimates by Flaw

Table 46 reports the POD estimates from logistic regression for each of the 68 round-robin flaws. Flaw and specimen characteristics are reported in the left-hand columns of the table, and three POD estimates are reported for each flaw in the three right-most columns. Each POD estimate was calculated using a different TPR. The column labeled “Mean TPR” reports an estimate of the POD calculated using the average of technician TPR over all 18 technicians participating in round-robin testing. The influence of technician skill on flaw detection rates was illustrated by estimating the POD for the technicians with the lowest TPR (i.e., technician 1) and the highest TPR (i.e., technician 17). The difference between the POD estimates based on the highest and lowest TPR is the range of the POD for each flaw within the pool of technicians who participated in round-robin testing.

Table 46. Estimates of the POD from logistic regression models for all 68 flaws scanned during round-robin testing.

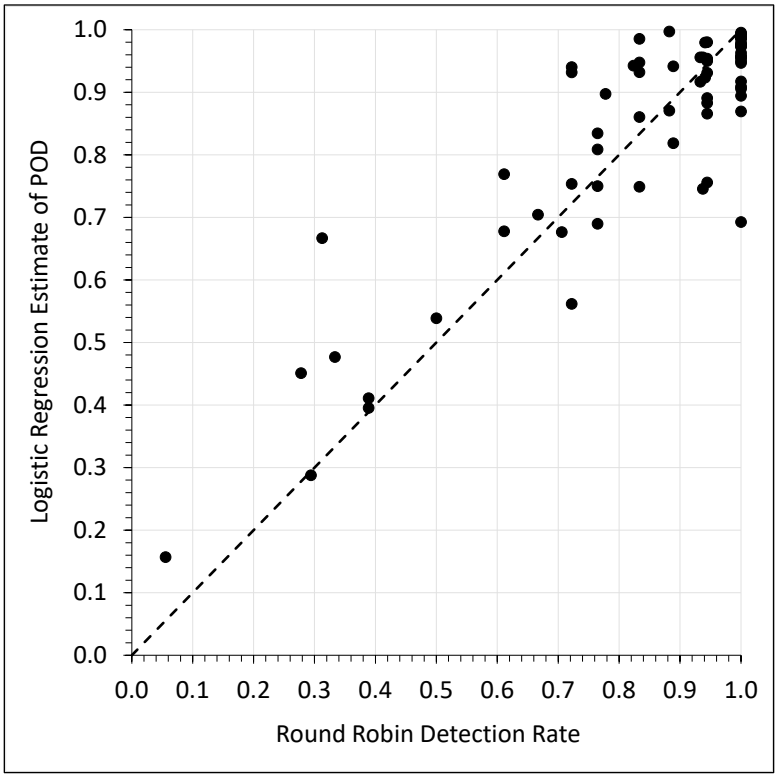
Flaw Category and Subcategory		Length (in.)	Height (in.)	Aspect	Area (in.²)	Joint Type	Joint Skew	Plate Thick. (in.)	Specimen Class	Technician TPR		
										Mean	Lowest	Highest
V	POR	2.77	0.14	0.0505	—	B	0	0.62	ERDC	0.1549	0.0717	0.2042
L	LAM	0.35	—	—	0.119	T	1	0.51	ERDC	0.3095	0.0064	0.6448
P	LOF	0.23	0.04	0.172	—	B	0	4	ERDC	0.3955	0.1544	0.5236
V	SLAG	6.98	0.11	0.0158	—	B	0	0.6875	ERDC	0.4028	0.2243	0.4900
P	TRCRK	0.37	0.18	0.486	—	C	0	1.87	ERDC	0.4509	0.1878	0.5819
P	ROCRK	0.08	0.07	0.875	—	B	0	0.64	ERDC	0.4767	0.2049	0.6080
P	BMCRK	0.3	0.13	0.424	—	C	0	1.375	ERDC	0.5389	0.2505	0.6679
P	TRCRK	0.16	0.16	1.000	—	B	0	0.5	ERDC	0.5616	0.2689	0.6889
P	ROCRK	0.41	0.1	0.244	—	B	0	0.76	AWS	0.6669	0.3731	0.7818
V	POR	0.98	0.21	0.2143	—	C	0	0.73	AWS	0.6759	0.4798	0.7539
P	LOF	0.39	0.11	0.282	—	T	1	0.73	AWS	0.6765	0.3801	0.7868
P	TOCRK	1.96	0.14	0.071	—	B	0	4	ERDC	0.6777	0.3820	0.7882
P	ROCRK	0.4	0.12	0.300	—	B	0	0.62	ERDC	0.6925	0.3991	0.7999
P	ROCRK	0.55	0.23	0.418	—	T	0	0.75	ERDC	0.7044	0.4133	0.8092
V	SLAG	2.37	0.05	0.0211	—	B	0	0.64	ERDC	0.7369	0.5529	0.8042
P	TRCRK	0.91	0.16	0.176	—	B	0	0.76	AWS	0.7457	0.4707	0.8426
P	TOCRK	0.19	0.11	0.579	—	T	1	0.73	AWS	0.7500	0.4723	0.8435
P	BMCRK	1.37	0.16	0.117	—	B	0	0.8125	ERDC	0.7536	0.4777	0.8463
P	ROCRK	1.27	0.15	0.118	—	B	0	0.625	ERDC	0.7558	0.4808	0.8479
P	TRCRK	1.05	0.14	0.133	—	B	0	0.625	ERDC	0.7691	0.4999	0.8575
L	LAM	0.71	—	—	0.5	B	0	0.64	ERDC	0.7838	0.0900	0.9652
P	LOF	1.47	0.24	0.164	—	T	1	0.51	ERDC	0.8088	0.5650	0.8866
P	BMCRK	1.47	0.32	0.218	—	T	0	0.75	ERDC	0.8186	0.5784	0.8920
V	POR	0.49	0.14	0.2857	—	T	1	0.73	AWS	0.8243	0.6763	0.8740
V	SLAG	0.56	0.21	0.3750	—	T	1	0.51	ERDC	0.8609	0.7363	0.9027
P	LOP	1.02	0.54	0.529	—	B	0	4	ERDC	0.8657	0.6650	0.9228
P	CLCRK	0.24	0.09	0.375	—	B	0	1	ERDC	0.8694	0.6723	0.9251
L	LAM	1.41	—	—	2	T	0	1.51	ERDC	0.8744	0.2052	0.9863
P	CLCRK	1.02	0.39	0.382	—	C	0	1.87	ERDC	0.8828	0.6997	0.9334
L	LAM	1.01	—	—	1.05	B	0	0.625	ERDC	0.8829	0.2251	0.9879
V	SLAG	1.09	0.13	0.1193	—	B	0	0.5	ERDC	0.8829	0.7717	0.9182
V	SLAG	0.28	0.08	0.2857	—	C	0	1.375	ERDC	0.8898	0.7838	0.9233
P	CLCRK	0.46	0.11	0.239	—	B	0	0.98	AWS	0.8944	0.7272	0.9413
P	LOF	0.46	0.13	0.283	—	T	0	0.98	AWS	0.9056	0.7523	0.9481

Table 46 (cont.). Estimates of the POD from logistic regression models for all 68 flaws scanned during round-robin testing.

Flaw Category and Subcategory		Length (in.)	Height (in.)	Aspect	Area (in.²)	Joint Type	Joint Skew	Plate Thick. (in.)	Specimen Class	Technician TPR		
										Mean	Lowest	Highest
P	CLCRK	2.59	0.38	0.147	—	C	0	1.87	ERDC	0.9088	0.7566	0.9493
P	LOF	0.64	0.13	0.203	—	B	0	1	AWS	0.9169	0.7686	0.9524
P	LOF	0.82	0.14	0.171	—	B	0	0.99	AWS	0.9174	0.7831	0.9560
P	LOF	0.27	—	—	—	B	0	0.75	AWS	0.9233	0.7913	0.9580
V	POR	0.28	0.2	0.7122	—	C	0	1.375	ERDC	0.9254	0.8483	0.9489
P	CLCRK	0.87	0.19	0.218	—	T	0	0.75	ERDC	0.9318	0.8110	0.9627
P	LOF	0.57	0.07	0.123	—	B	0	0.625	ERDC	0.9320	0.8116	0.9629
P	CLCRK	0.85	0.45	0.530	—	B	0	1	ERDC	0.9404	0.8324	0.9676
P	LOF	0.86	0.23	0.267	—	C	0	0.73	AWS	0.9426	0.8399	0.9693
P	LOF	2.11	0.2	0.095	—	B	0	0.8125	ERDC	0.9468	0.8490	0.9713
P	CLCRK	1.27	0.34	0.268	—	B	0	0.8125	ERDC	0.9499	0.8569	0.9730
P	CLCRK	1.72	0.25	0.145	—	B	0	0.6875	ERDC	0.9500	0.8573	0.9731
P	LOF	4.52	0.21	0.046	—	T	0	0.75	ERDC	0.9539	0.8674	0.9752
P	TOCRK	0.54	0.14	0.261	—	B	0	0.99	AWS	0.9549	0.8745	0.9767
P	TOCRK	0.46	0.25	0.543	—	B	0	1	AWS	0.9557	0.8681	0.9754
P	TOCRK	0.59	0.09	0.153	—	T	0	0.98	AWS	0.9560	0.8749	0.9768
P	TOCRK	0.76	0.34	0.448	—	C	0	1.375	ERDC	0.9570	0.8758	0.9770
P	TOCRK	0.61	0.12	0.197	—	B	0	0.98	AWS	0.9574	0.8783	0.9775
V	SLAG	0.77	0.08	0.1039	—	T	0	0.98	AWS	0.9594	0.9152	0.9729
V	POR	1.37	0.26	0.1898	—	B	0	4	ERDC	0.9719	0.9401	0.9812
V	POR	0.68	0.17	0.2500	—	B	0	0.76	AWS	0.9722	0.9413	0.9816
P	LOP	0.65	0.04	0.068	—	B	0	0.99	AWS	0.9727	0.9219	0.9861
P	LOP	0.57	0.19	0.333	—	B	0	1	ERDC	0.9759	0.9280	0.9873
P	LOP	0.29	0.1	0.345	—	B	0	0.5	ERDC	0.9793	0.9378	0.9891
P	LOP	0.67	0.09	0.139	—	B	0	0.75	AWS	0.9797	0.9393	0.9894
P	TOCRK	3.35	0.51	0.152	—	T	0	1.51	ERDC	0.9800	0.9399	0.9895
P	LOP	1.31	0.16	0.122	—	B	0	0.64	ERDC	0.9854	0.9557	0.9924
P	LOP	1.47	0.16	0.109	—	B	0	0.62	ERDC	0.9859	0.9572	0.9926
P	LOP	3	0.16	0.053	—	B	0	0.64	ERDC	0.9867	0.9595	0.9930
P	LOP	2.79	0.16	0.057	—	B	0	0.62	ERDC	0.9868	0.9600	0.9931
V	SLAG	0.62	0.21	0.3387	—	C	0	0.73	AWS	0.9903	0.9790	0.9936
V	SLAG	1.16	0.1	0.0836	—	B	0	0.75	AWS	0.9931	0.9849	0.9954
V	SLAG	1.29	0.12	0.0930	—	B	0	1	AWS	0.9951	0.9889	0.9966
V	SLAG	0.6	0.08	0.1333	—	B	0	0.98	AWS	0.9969	0.9933	0.9980

In Figure 30, estimates of POD are compared to actual detection rates for each flaw. The actual detection rate is the fraction of times a flaw was scanned and detected. The dashed diagonal line is the line of perfect agreement between the POD estimates and the actual detection rates. POD estimates above the line are greater than the actual detection rates, and those below the line are lower than the actual detection rates. For those flaws that were frequently missed, there appears to be a tendency to overestimate POD. The correlation between the POD estimates and actual detection rates is $r = 0.8859$.

Figure 30. Estimated POD and observed round-robin detection rate for each flaw.



4.6 Reliability of Reported Indications

The reliability of a reported indication is the probability that an indication represents an actual flaw (i.e., a TP) and is not a false call (i.e., a FP). When a discontinuity is reported in the field, information about its dimensions and other characteristics are also reported by the technician. In this section, the reliability of reported indications is modeled to assess the degree to which the characteristics reported by a technician might be more or less indicative of false calls than others. If so, these models may be of some practical importance in terms of identifying red flags when evaluating

technician reports. Red flags might suggest that additional review is needed to verify a reported indication is an actual flaw.

A logistic regression was fit to all indications reported by technicians during round-robin testing. Each indication was assigned a value of 1 if it was classified as a TP indication and a value of 0 if it was classified as an FP indication. TP indications were those that corresponded to a known flaw in a weld specimen. FP indications were false calls. Independent variables included the flaw dimensions and characteristics reported by the technician and observable characteristics of the specimen, including joint type, joint skew, and plate thickness. Table 47 summarizes the results. Significant continuous independent variables included the length and aspect of the discontinuity reported by the technician and the technician's round-robin PPV. In terms of categorical variables, flaw category, joint type, and specimen class were significant. Joint skew was not significant. Of the 1,112 indications used to estimate this logistic regression model, 981 indications were classified as TP indications, and 131 were classified as FP indications.

The reliability of a nominal indication is calculated by setting each continuous independent variable at its mean and each categorical variable at 0 (zero). Table 48 summarizes the mean values used to define the nominal indication. Table 49 summarizes the effects of observable specimen characteristics on the reliability of reported indications. The reliability of the nominal indication was 0.9075, which means that, in the absence of information about the indication, the specimen, or the technician, there is roughly a 1 in 10 chance that the indication is false. In evaluating the reliability of an indication, the single most important factor to consider is technician PPV. The reliability of indications was 0.7683 for the technician with the lowest round-robin PPV and 0.9465 for the technician with the highest round-robin PPV. In other words, nearly 1 in 4 of the indications reported by the technician with the lowest PPV were false calls. The technician with the lowest PPV was five times more likely to report a false indication than the technician with the highest PPV. Indications in butt joints had a higher reliability than indications in T- and corner joints. Indications that were reported to be LOF or unknown (UNK) were less reliable than those classified as other flaw categories. Indications reported as LOP, on average, had the highest reliability (0.9487). Indications in ERDC specimens were 67.5% more likely to be false than indications in AWS specimens. The probability of a false indication in ERDC specimens was 0.128, while the same probability in AWS specimens was 0.0764.

Table 47. Logistic regression results for the reliability of an indication.

Model Version		Initial Model				Revised Model			
Parameter	Discrete Level	Param. Estimate	Std. Error	Wald Chi-Square	Pr > Chi-Square	Param. Estimate	Std. Error	Wald Chi-Square	Pr > Chi-Square
Intercept	—	−1.3806	0.9948	1.926	0.1652	−1.6985	0.7164	5.6207	0.0177
Length reported	—	0.3997	0.237	2.8453	0.0916	0.6535	0.2107	9.6233	0.0019
Height reported	—	1.9836	1.3662	2.1081	0.1465	—	—	—	—
Aspect reported	—	−2.2336	0.6737	10.9911	0.0009	−1.5305	0.4513	11.4986	0.0007
Plate thickness reported	—	0.0438	0.1604	0.0744	0.785	—	—	—	—
Technician PPV	—	3.8248	1.0858	12.4079	0.0004	4.3392	0.7977	29.5862	<.0001
Omniscan instrument (Ref = X3)	MX2	0.0358	0.1955	0.0335	0.8547	—	—	—	—
	MX3	−0.2271	0.3543	0.4107	0.5216	—	—	—	—
	SX	0.2847	0.254	1.2571	0.2622	—	—	—	—
Joint type (Ref = T)	B	0.3161	0.1536	4.2349	0.0396	0.3644	0.1384	6.931	0.0085
	C	−0.1051	0.1966	0.2856	0.5931	0.0111	0.1811	0.0037	0.9512
Joint skew (Ref = TRUE)	FALSE	0.2073	0.199	1.0853	0.2975	—	—	—	—
Flaw category and subcategory reported (Ref = V_SLAG)	L_LAM	0.1706	0.5013	0.1158	0.7336	0.0296	0.4855	0.0037	0.9514
	P_CRK	0.1216	0.2176	0.3123	0.5762	0.2153	0.2082	1.0696	0.301
	P_LOF	−0.6045	0.2142	7.9673	0.0048	−0.5816	0.2119	7.5298	0.0061
	P_LOP	0.6324	0.3578	3.1236	0.0772	0.7108	0.3558	3.9902	0.0458
	V_POR	−0.4027	0.4552	0.7826	0.3763	0.1494	0.4004	0.1392	0.7091
	UNK	0.175	0.404	0.1878	0.6648	−0.4399	0.4408	0.9961	0.3183
Class (Ref = ERDC)	AWS	3.8248	1.0858	12.4079	0.0004	0.2869	0.1132	6.423	0.0113
Model fit statistics		Likelihood ratio chi-square = 129.4932 ($p < 0.00001$, $df = 18$)				Likelihood ratio chi-square = 124.1170 ($p < 0.00001$, $df = 12$)			
		AIC, intercept: 808.267 AIC, intercept, and covariates: 714.824				AIC, intercept: 808.267 AIC, intercept, and covariates: 708.1500			
		Percent concordant: 77.3% Area under the ROC curve: 0.7725				Percent concordant: 76.8% Area under the ROC curve: 0.7682			
Response variable		1 = TP indication; 0 = FP indication							
No. of observations		n = 1,112 (1 = 981, 0 = 131)							

Table 48. Descriptive statistics for continuous independent variables used to estimate the nominal reliability of reported indications.

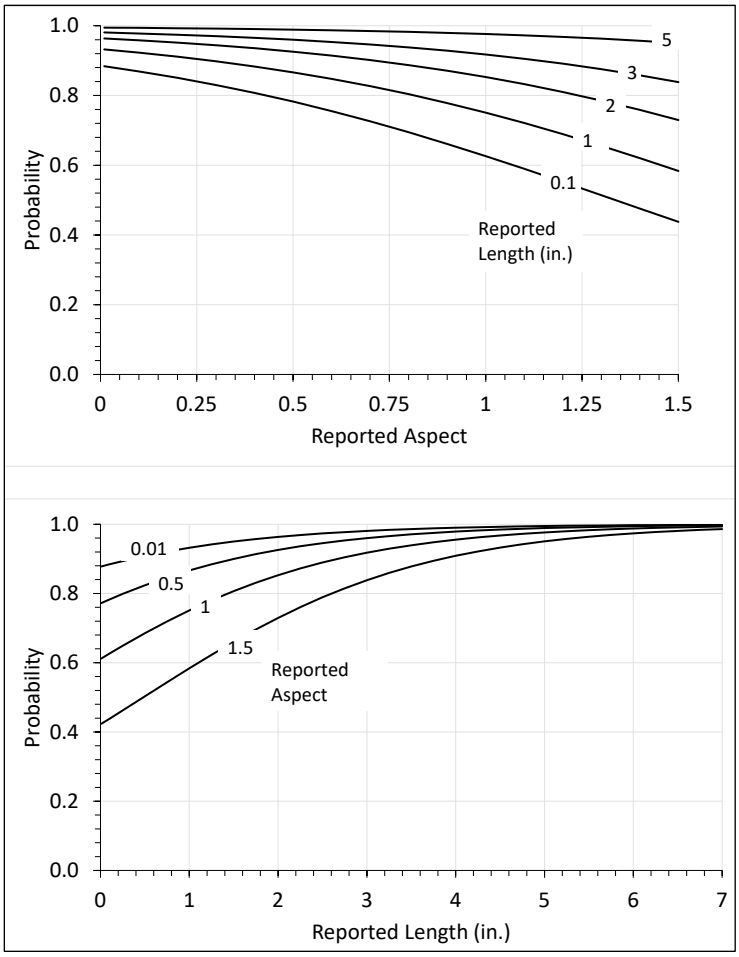
Independent Variables	Statistics				Number of Observations
	Mean	Std. Dev.	Min.	Max.	
Length reported (in.)	1.0455	0.9210	0.0000	7.8000	1118
Height reported (in.)	0.1810	0.1278	0.0100	0.8000	1118
Aspect reported	0.2482	0.2249	0.0042	3.2100	1112
Plate thickness (in.)	1.0380	0.7511	0.5000	4.0000	1157
Technician PPV	0.8479	0.1127	0.5978	0.9836	1157

Table 49. Estimates of the probability that a reported indication is an actual flaw.

Factor	Factor Level	Probability Indication is a Flaw
—	Nominal	0.9075
Technician PPV	Lowest PPV	0.7683
	Highest PPV	0.9465
Joint type	Butt joint	0.9289
	Corner joint	0.9018
	T-joint	0.8618
Flaw category	SLAG	0.8931
	LAM	0.9034
	CRK	0.9184
	LOF	0.8354
	LOP	0.9487
	UNK	0.8540
	POR	0.9134
Specimen class	ERDC	0.8720
	AWS	0.9236

Table 49 shows the sensitivity of the nominal flaw to differences in technician PPV. Figure 31 illustrates the sensitivity of nominal indication reliability to reported length and aspect. The aspect reported by a technician was calculated from the height and length reported by that NDT technician (Table 48). The greater the aspect of the reported flaw, the lower the reliability of the indication for a given length. However, this effect diminishes as the reported length of the indication increases. The reliability of reported indications increases with reported length (Figure 31), as might be expected.

Figure 31. Reliability of reported indications as a function of reported flaw length and aspect.

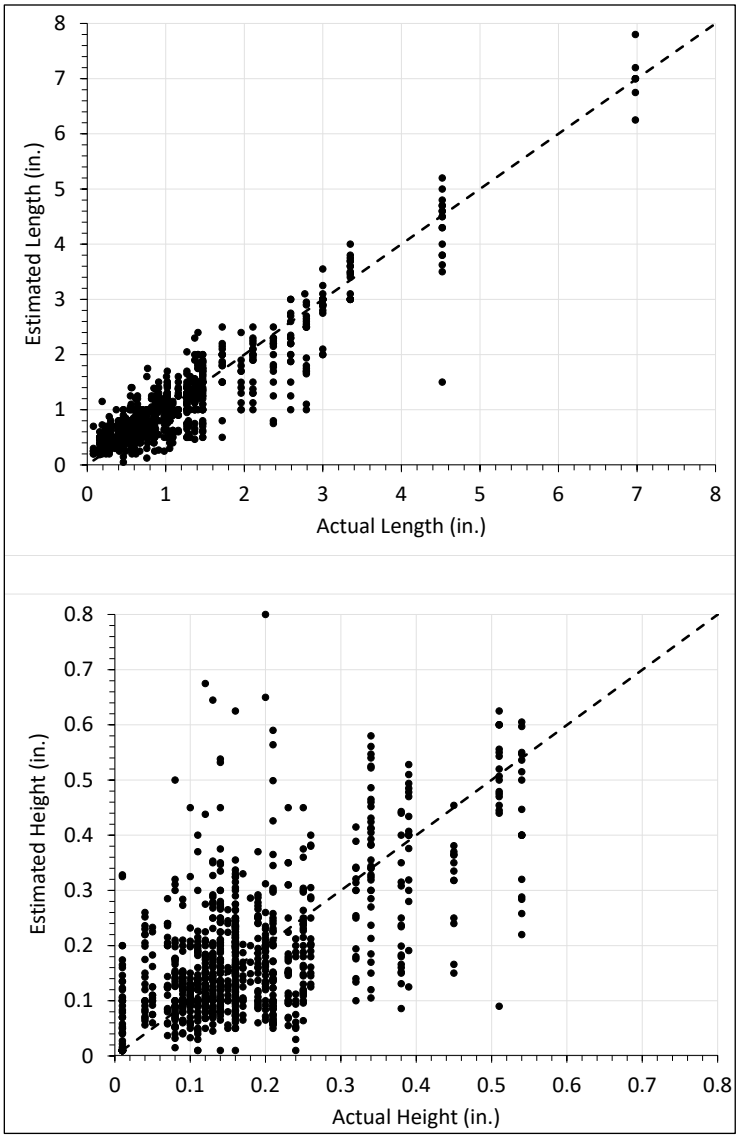


4.7 Uncertainty in Estimates of Flaw Length and Height

NDT technicians estimated the length and height of detected flaws using the decibel drop methods and diffraction techniques. In Figure 32, estimates of length and height are plotted against the actual length and height. More scatter along the line of perfect agreement indicates greater uncertainty in estimates of flaw length and height. Actual flaw lengths ranged from 0.08 to 6.98 in. Points on the dashed diagonal line of perfect agreement represent estimates of flaw length containing little or no error. Points above the line represent overestimates of flaw length, and points below the line represent underestimates of flaw length. Most points fall below the line, suggesting a tendency to underestimate flaw length. Despite this, there is a high degree of linear association between estimated and actual flaw lengths. The Pearson correlation coefficient for estimated and actual length was 0.9304 ($p < 0.0001$). The correlation between actual and

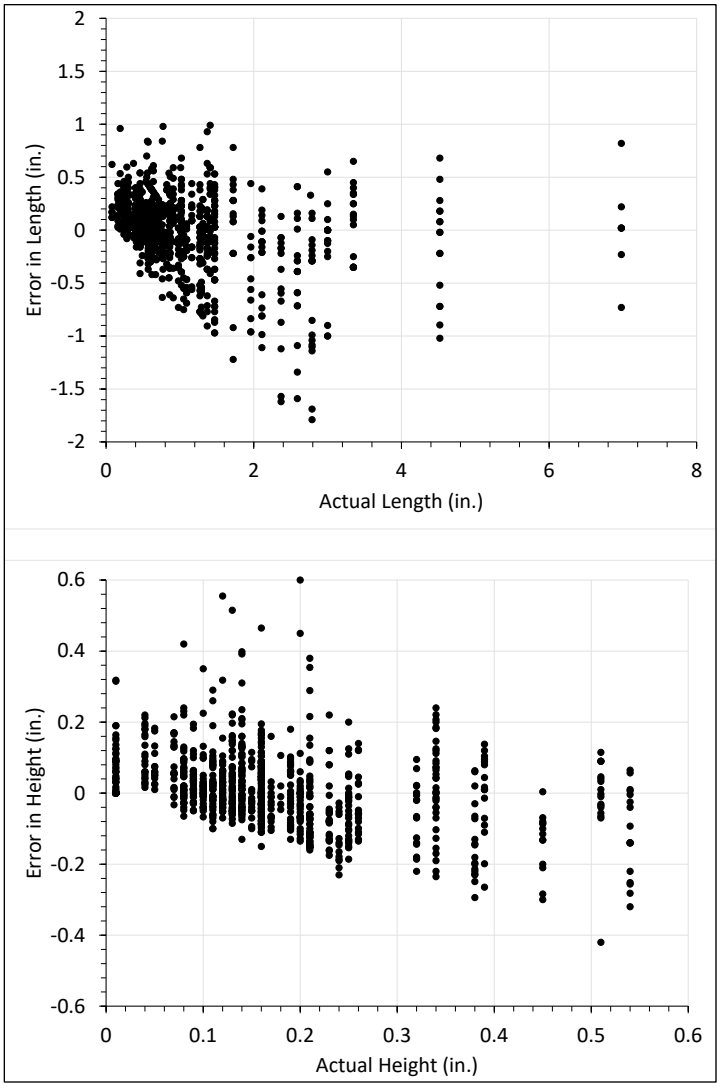
estimated flaw height shows more uncertainty in estimates of height than estimates of length. The actual heights of nonlaminar flaws ranged from 0.04 to 0.54 in. Four laminar flaws, which are not associated with a height, and one flaw classified as LOF were assigned a height of 0.01 for comparison with the technicians' height estimates. Figure 32 shows proportionally more error in estimates of height than length, and this is reflected in a much lower degree of linear association between estimated and actual height. The correlation between estimated and actual height was 0.6143 ($p < 0.0001$). Level II NDT technicians tended to overestimate flaw heights for flaws with heights less than 0.3 in. and to underestimate the flaw heights for flaws greater than 0.3 in. This result may have implications when using NDT estimates of flaw height in FFS analysis.

Figure 32. Actual and estimated flaw length and height.



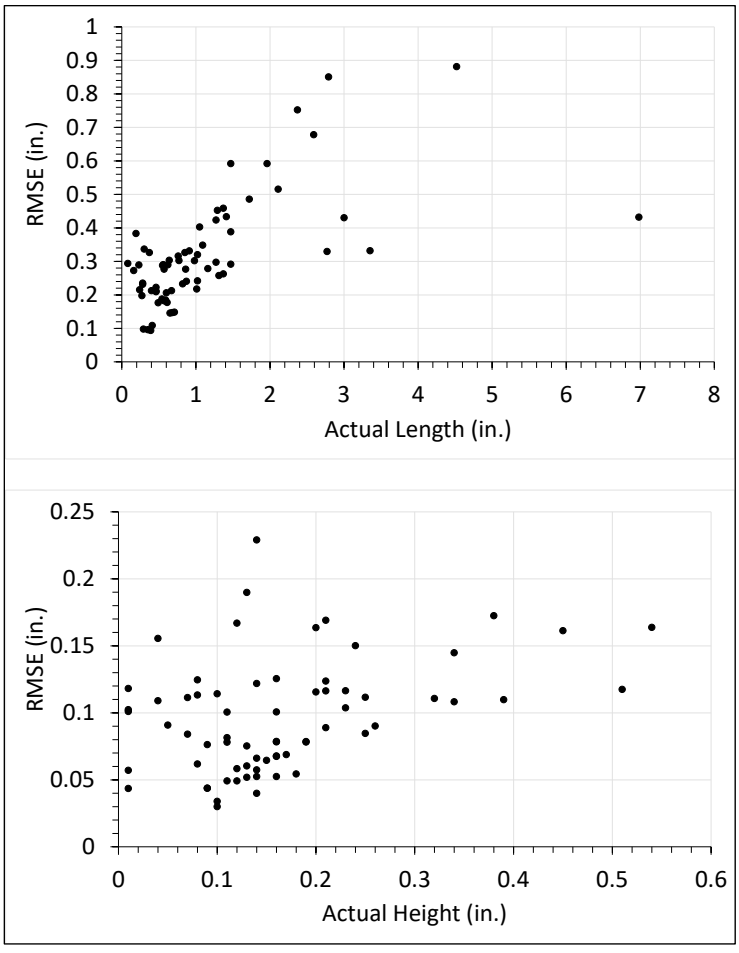
The error in an estimate of flaw length or height is the difference between the estimated and actual value. Figure 33 shows the errors in estimates of flaw length and height. Points above the line $y = 0$ represent overestimates, and points below the line represent underestimates. Most errors in length were in the range of +1 to −2 in., but there was one outlier point (4.25, −3.02) that is not shown in the figure. There was a tendency to overestimate the length of the shorter flaws and to underestimate the length of the longer flaws, but the errors for the longest flaws (i.e., ≥3 in.) appear more or less evenly distributed around the $y = 0$ line. Errors in estimates of flaw height also appeared to increase with actual flaw height. Technicians always overestimated the height of flaws that were less than 0.05 in. There was a strong tendency to underestimate the height of flaws that were greater than about 0.2 in. Figure 33 displays this tendency.

Figure 33. Error in estimates of flaw length and height.



The RMSE is the average difference between a set of estimated and actual values. RMSE is a straightforward measure of how well the estimates match the data. For the flaws as a whole, the RMSE for length was 0.363 in. (9.23 mm), and the RMSE for height was 0.106 in. (2.70 mm). However, a single value of RMSE for all flaws did not reflect the tendency for larger flaws to be associated with larger errors in flaw size estimates. Figure 34 plots the RMSE for each flaw to show that, particularly with respect to length, there was a tendency for larger errors to be associated with larger flaws. This observation led to an analysis of proportional error in length and height estimates.

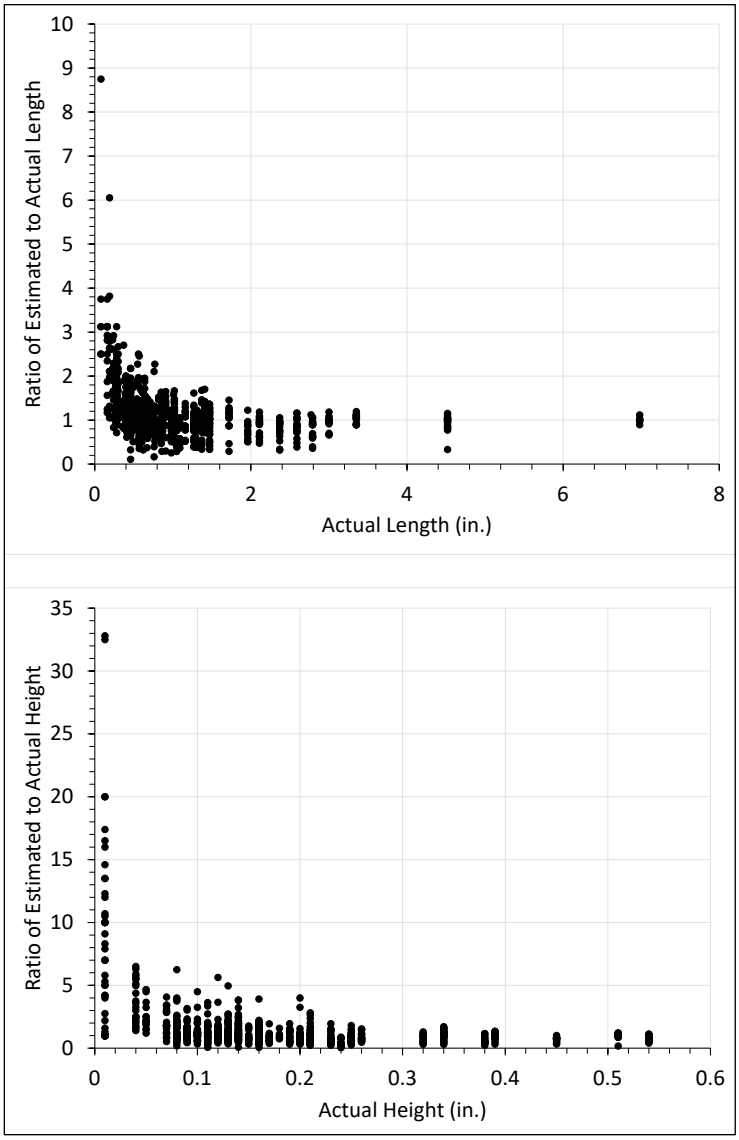
Figure 34. Root-mean-square error (RMSE) by actual flaw length and height.



Errors in estimates of flaw length and height were analyzed in proportional terms as the ratio of the estimated length or height to the actual length or height. A ratio of one means the estimate contains no error. Estimates for which the ratio is less than one are underestimates, and

estimates for which the ratios are greater than one are overestimates. For example, a ratio of 0.5 for length indicates that a flaw is twice as long as the estimate reported by the technician. A ratio of 2 indicates that the flaw is half as long as reported by the technician. The lower bound on the ratio is 0 (zero). Figure 35 shows the ratios of estimated to actual length and height. The largest ratios are associated with the smaller flaws. Note the differences in scale on the *x*- and *y*-axes of the two plots in Figure 35.

Figure 35. Ratios of estimated to actual flaw length and height.

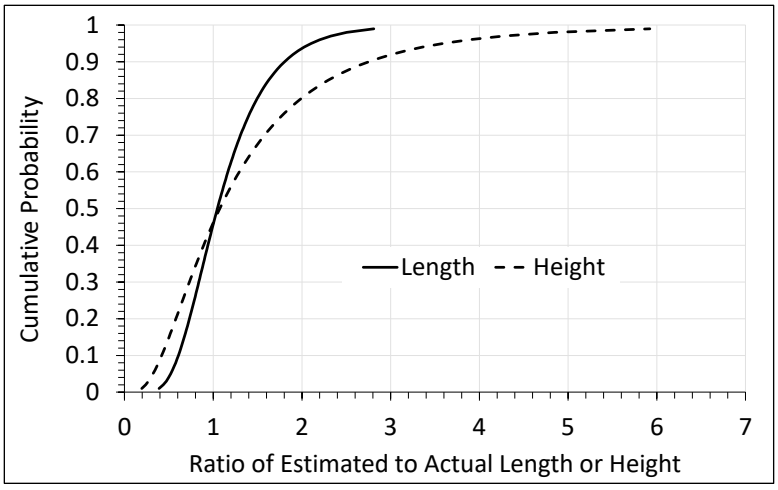


Flaws tended to have greater length than height, and errors in estimates of height were proportionally greater than errors in estimates of length. The flaws with the smallest actual height in Figure 35 are the five flaws that were assigned a height of 0.01 in. for analysis. Although the error in these

estimates of height may be small in absolute terms (Figure 33), they are large in proportional terms. In particular, technicians consistently overestimated the height of laminar flaws. While height is not typically reported for laminar flaws, these technicians reported heights for laminar flaws approximately 40% of the time. This was sometimes, but not always, accompanied by a misclassification of flaw category. Estimated heights were as much as 33 times greater than the assigned height of 0.01 in. For the group of flaws with the next greatest actual height, 0.04 in., the ratios of estimated to actual flaw height did not exceed seven.

Figure 36 shows the variability in the ratios of estimated to actual flaw length and height using lognormal distributions. These lognormal distributions describe uncertainty in the proportional error of flaw length and height estimates. The lower and upper bounds on the 90% confidence interval for the ratio of estimated to actual length are 0.5205 and 2.1041, respectively (Table 50, last row). The lower and upper bounds on the confidence interval for ratio of estimated to actual height are 0.3212 and 3.5911, respectively (Table 50). These results indicate that errors in estimating length vary by a factor of two, and errors in estimating height vary by a factor of more than three.

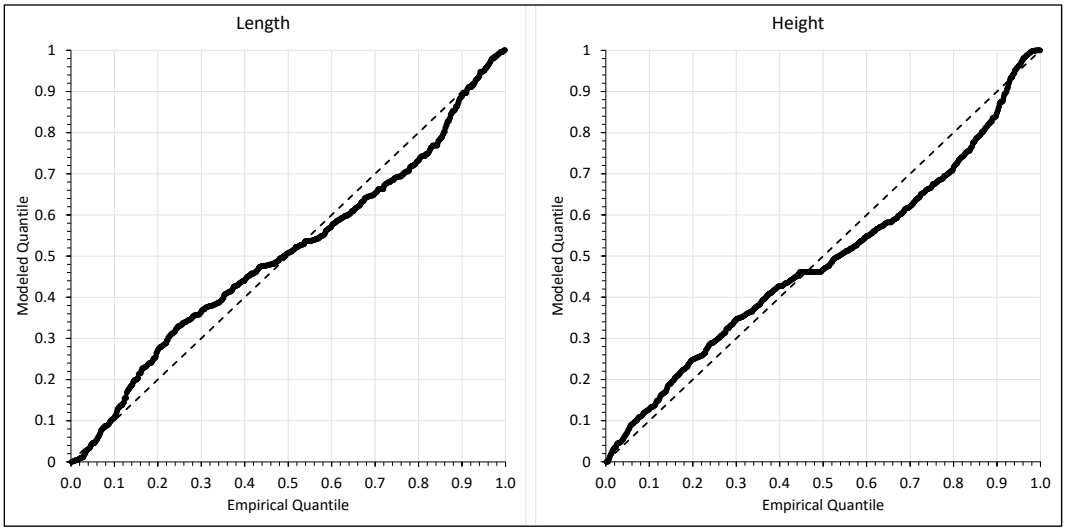
Figure 36. Cumulative distribution functions describing uncertainty in length and height estimates.



The lognormal distributions are models of the round-robin testing data. The lognormal distribution is often chosen to model the distribution of ratios, and the quantile–quantile (Q-Q) plots in Figure 37 demonstrate the appropriateness of this distribution. In Q-Q plots, the theoretical and observed quantiles of each data point are plotted in a Cartesian plane to

provide a visual indication of how closely the distribution fits the data. The distribution fits the data if all of the points fall along the dashed line of perfect agreement. The tails of a distribution are often difficult to fit, and one particular advantage of Q-Q plots is they make it easy to see how well the distribution fits in the tails. Figure 37 shows the lognormal distribution fits these data reasonably well in both the body and the tails.

Figure 37. Quantile–quantile (Q-Q) plots for lognormal distributions fit to the ratio of estimated to actual length and height.



These characterizations of uncertainty are obtained by lumping all round-robin scans into a single group. It is possible that subsets of scans by technician or by flaw and joint type could contain more or less error. For example, there may be less error in estimates by some technicians than others because they are more skillful. Similarly, there may be more error in specimens with T- or corner joints because access to the weld is more limited. The subsequent sections of this report explore whether certain flaw and specimen characteristics are associated with more or less error than others.

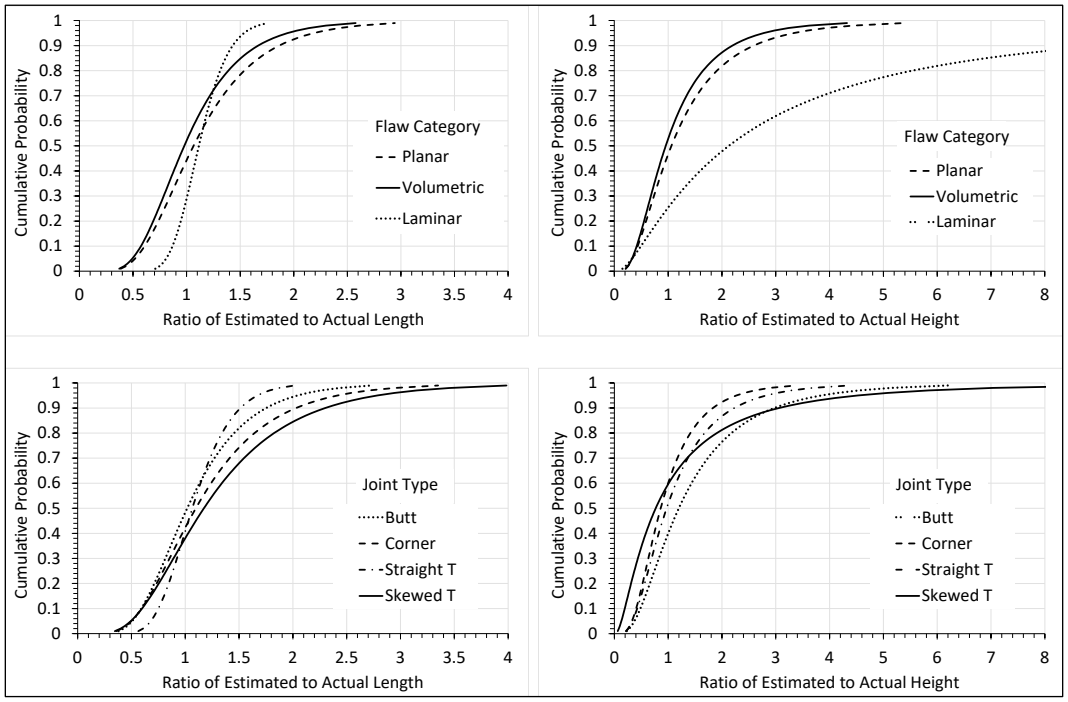
4.7.1 Uncertainty by Flaw and Joint Category

Table 50 summarizes the uncertainty in the ratio of estimated to actual length and height by flaw and joint type. Lognormal distributions were used to fit the data. Table 50 reports the mean and standard deviations of the log-transformed ratios, which are the parameters of the lognormal distribution. The columns labeled $F_{0.05}$ and $F_{0.95}$ are the lower and upper 90% confidence bounds on each ratio, respectively. Figure 38 plots the lognormal distributions for ratios of estimated to actual length and height.

Table 50. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height by flaw and joint type and for all flaws.

Subsets of Length and Height Estimates		Ratio of Estimated to Actual Length				Ratio of Estimated to Actual Height			
		$\mu_{\ln(l)}$	$\sigma_{\ln(l)}$	$F_{0.05}$	$F_{0.95}$	$\mu_{\ln(h)}$	$\sigma_{\ln(h)}$	$F_{0.05}$	$F_{0.95}$
Flaw category	Planar	0.0625	0.4371	0.5187	2.1850	0.0597	0.6966	0.3375	3.3385
	Volumetric	−0.0221	0.4165	0.4930	1.9407	−0.0461	0.6493	0.3282	2.7785
	Laminar	0.1083	0.1947	0.8089	1.5352	0.7544	1.1375	0.3274	13.8116
Joint type	Butt	0.0177	0.4237	0.5070	2.0435	0.1809	0.7089	0.3734	3.8456
	Corner	0.0935	0.4794	0.4991	2.4160	−0.1497	0.5889	0.3268	2.2679
	Straight T	0.0634	0.2738	0.6791	1.6718	−0.0287	0.6482	0.3346	2.8221
	Skewed T	0.1597	0.5256	0.4942	2.7850	−0.2632	1.0780	0.1305	4.5267
All estimates		0.0454	0.0714	0.5205	2.1041	0.4246	0.7338	0.3212	3.5911

Figure 38. Uncertainty in the ratio of estimated to actual length and height by flaw category and joint type.



Figures 39 and 40 show the Q-Q plots for each flaw and joint type to provide a visual check on how well the distribution fits and to demonstrate that the lognormal distribution is a good model for the proportional error. The fit is not as good for laminar flaw height because height is undefined for laminations, which leaves a large portion of the data clustered in the lower tail. In these cases, technicians correctly classified the laminations and assigned a value for height that was less than or equal to the nominally assigned height (0.01 in.). Technicians assigned a height greater than 0.01 in. to 19 of the 51 laminations that were detected. In 8 of those 19 cases, the technician misclassified the flaw category. However, in 11 of

those 19 cases, the technician estimated a height greater than 0.01 in. despite recognizing that the flaw was a lamination.

The 90% confidence interval for ratio of estimated to actual length was similar across flaw categories, but there was less uncertainty associated with length estimates for laminar flaws. The variability in the ratio of estimated to actual flaw height was similar for planar and volumetric flaws. The variability in the ratio of estimated to actual height was much greater for laminar flaws, which were assigned a nominal height of 0.01. The variability in the ratio of estimated to actual length and height was similar across joint types. The slightly higher upper bound on variability in the ratio for flaws in butt joints indicates a tendency to overestimate the height of flaws in butt joints. This was contrary to expectations because butt joints offer greater accessibility than T- and corner joints, but it may be attributed to the larger number of butt joints represented among the round-robin specimens.

Figure 39. Q-Q plots for lognormal distributions fit to the ratio of estimated to actual length and height by flaw category.

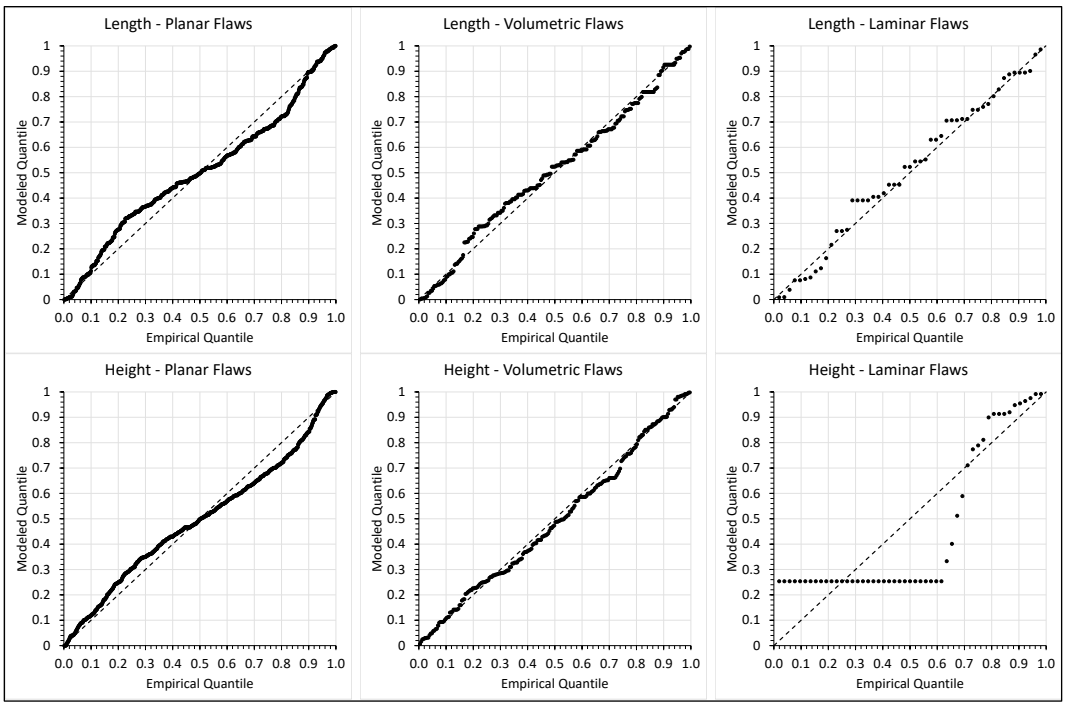
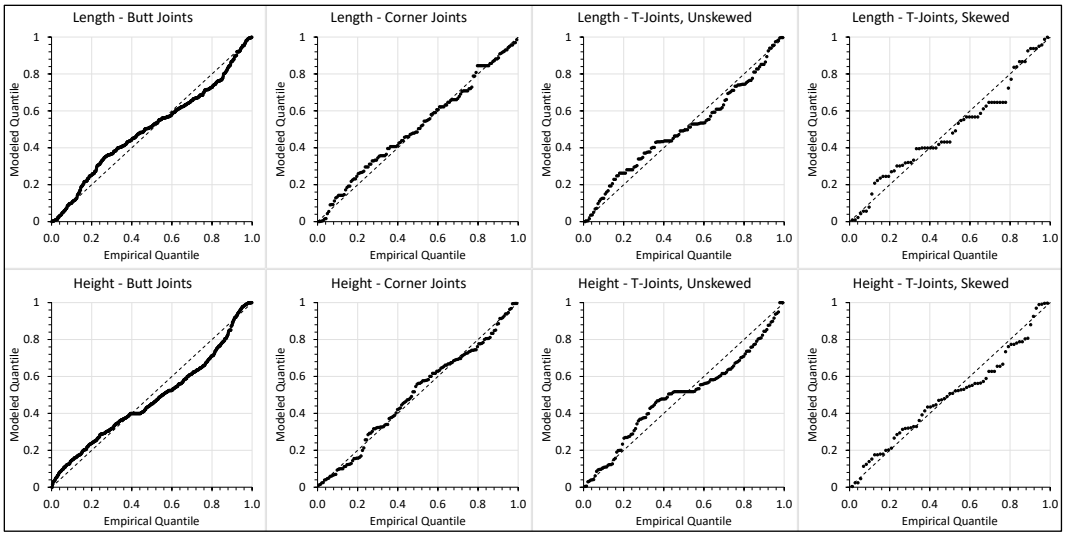


Figure 40. Q-Q plots for lognormal distributions fit to the ratio of estimated to actual length and height by joint type.



4.7.2 Uncertainty by Flaw Subcategory

Table 51 summarizes the uncertainty in the ratio of estimated to actual length and height by flaw subcategory. The objective here is to explore whether or not there may be more or less error associated with some flaw subcategories than others. For the ratio of estimated to actual length, the lower bounds of the 90% confidence interval, $F_{0.05}$, ranged from 0.43 to 0.81, and the upper bound, $F_{0.95}$, ranged from 1.54 to 2.96. TRCRK was the flaw subcategory associated with the greatest uncertainty in length. For height, the lower bounds for the 90% confidence interval for height ranged from 0.28 to 0.56. The upper bound ranged from 1.61 to 6.60 (excluding laminations). The largest uncertainties in height were associated with the flaw category LOF. Figure 41 illustrates the distributions for flaw length and height. The Q-Q plots in Figures 42 and 43, respectively, demonstrate the fit of lognormal distribution to length and height data for each flaw subcategory. These plots demonstrate that the lognormal distribution works consistently well as a model of the proportional error in subsets of the round-robin data. The Q-Q plots for laminar length and height are identical to those presented for the laminar flaw category in Figure 39 because that category has only one subcategory.

Table 51. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height by flaw category.

Flaw Category	Flaw Subcategory	Ratio of Estimated to Actual Length (l)				Ratio of Estimated to Actual Height (h)			
		$\mu_{\ln(l)}$	$\sigma_{\ln(l)}$	$F_{0.05}$	$F_{0.95}$	$\mu_{\ln(h)}$	$\sigma_{\ln(h)}$	$F_{0.05}$	$F_{0.95}$
Planar	BMCRK	0.1259	0.3926	0.59	2.16	−0.1134	0.6273	0.32	2.51
	CLCRK	0.0116	0.4306	0.50	2.05	−0.1377	0.5373	0.36	2.11
	ROCRK	0.1602	0.5630	0.46	2.96	−0.1315	0.5140	0.38	2.04
	TOCRK	0.1335	0.4387	0.56	2.35	0.1202	0.4615	0.53	2.41
	TRCRK	0.2100	0.6357	0.43	3.51	0.0451	0.5854	0.40	2.74
	LOF	0.0598	0.4073	0.54	2.07	0.2549	1.0358	0.23	7.09
	LOP	−0.0415	0.3200	0.57	1.62	0.0987	0.5727	0.43	2.83
Volumetric	POR	0.0492	0.4170	0.53	2.09	−0.2888	0.4606	0.35	1.60
	SLAG	−0.0597	0.4127	0.48	1.86	0.0818	0.6974	0.34	3.42
Laminar	LAM	0.1083	0.1948	0.81	1.54	0.7544	1.1375	0.33	13.81

Figure 41. Cumulative distribution functions showing uncertainty in the ratio of estimated to actual length and height by flaw subcategory.

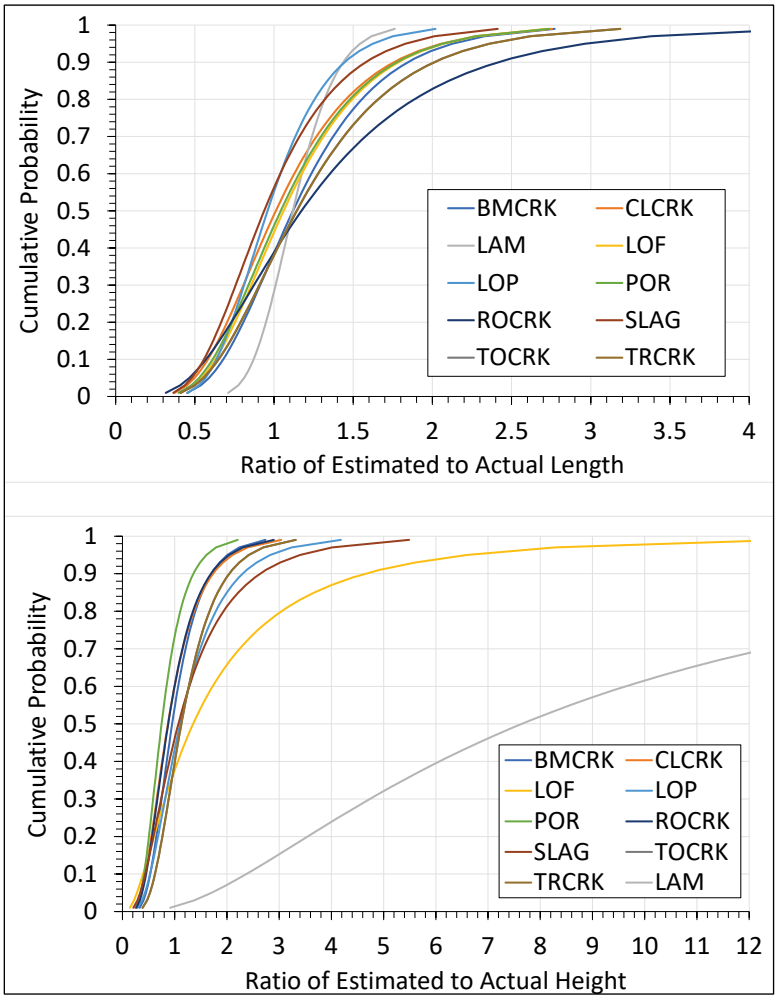


Figure 42. Q-Q plots showing the fit of lognormal distributions to the ratio of estimated-to-actual length by flaw subcategory.

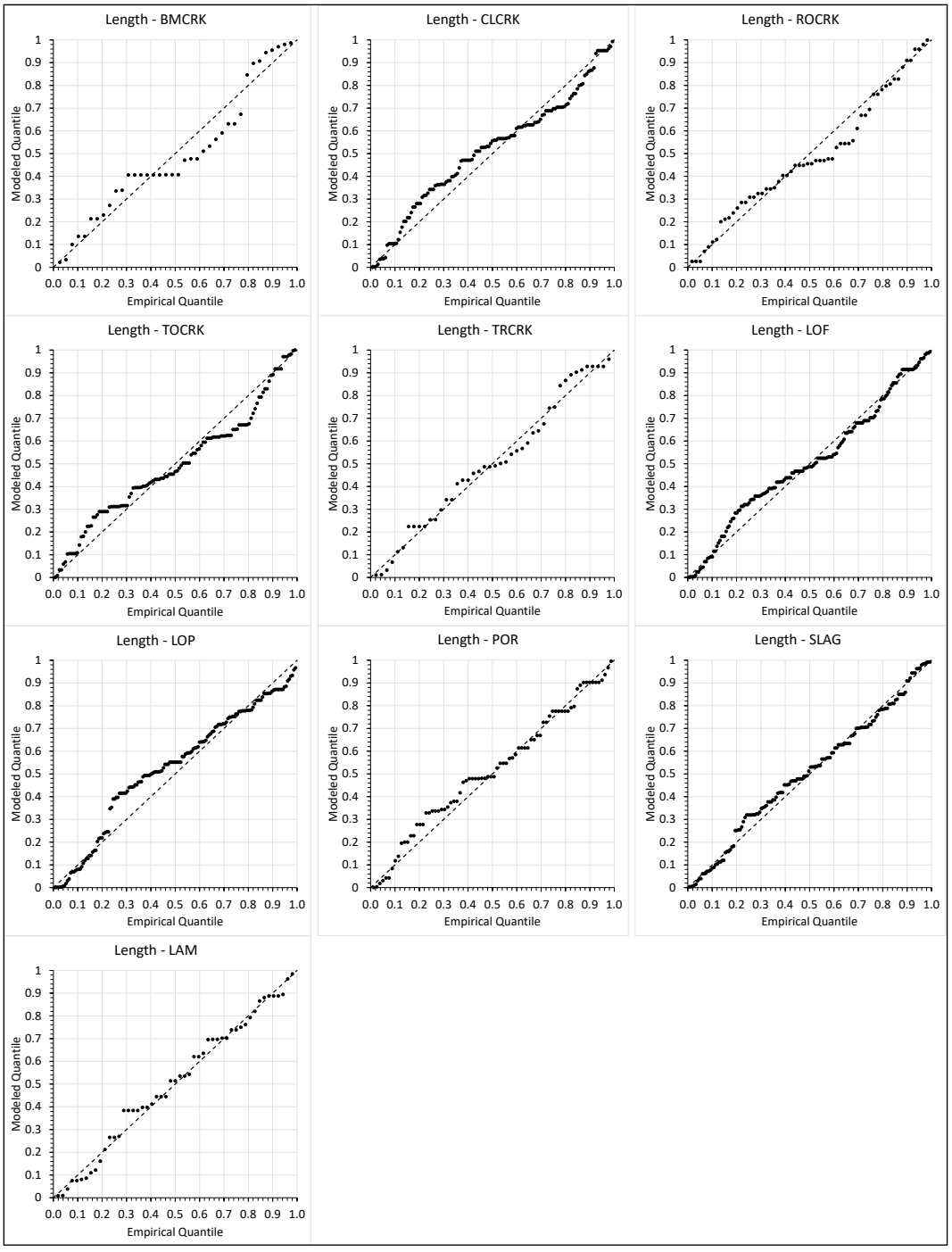
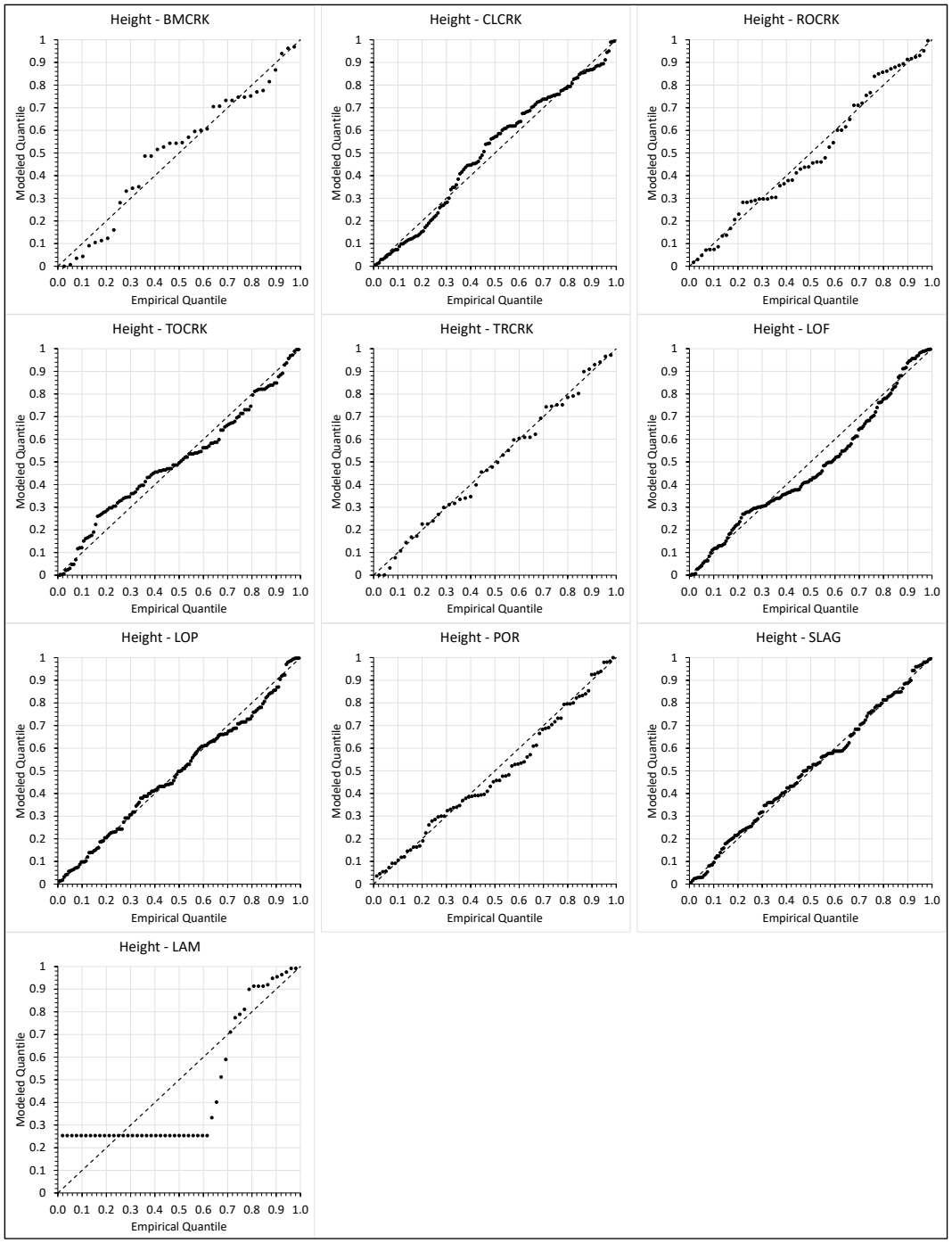


Figure 43. Q-Q plots showing the fit of lognormal distributions to the ratio of estimated to actual height by flaw subcategory.



4.7.3 Uncertainty by Technician

Uncertainty in the ratio of estimated to actual length and height varied by technician. Table 52 summarizes this variability. Lognormal distributions are used to characterize uncertainty in the proportional error, so the mean

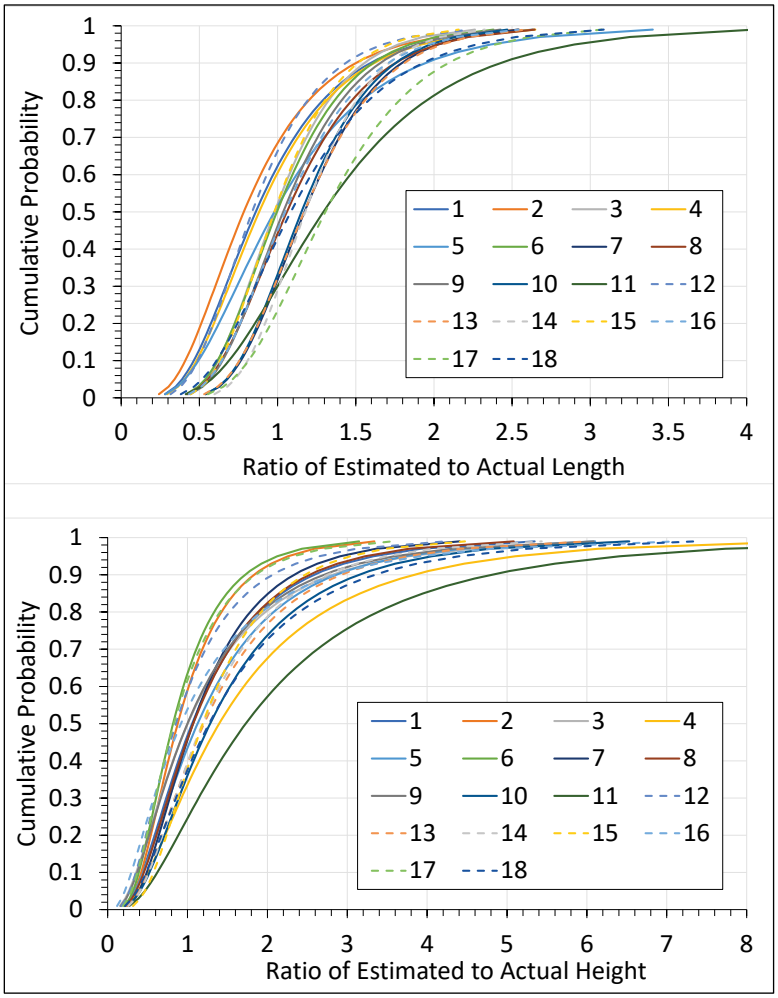
and standard deviation of the log-transformed ratios are listed along with each distribution's lower and upper 90% confidence bounds: $F_{0.05}$ and $F_{0.95}$, respectively. The results showed that the technicians varied considerably in their ability to accurately size flaws. The lower bounds of the 90% confidence interval on the ratio of estimated to actual length ranged from 0.34 to 0.71. The upper bounds ranged from 1.67 to 2.9. The lower bounds on the 90% confidence interval on the ratio of estimated to actual height ranged from 0.22 to 0.47. The upper bounds ranged from 2.12 to 6.41. These results can be compared to those obtained by the Level III NDT technician who conducted verification testing using PE. The 90% confidence bounds characterizing uncertainty in that technician's length and height estimates are summarized in the last row of Table 52. There is less uncertainty in these estimates of length and height than in those of the Level II technicians who applied PAUT.

Table 52. Parameters and confidence bounds for distributions on the ratio of estimated to actual flaw length and height.

Technician ID	Ratio of Estimated to Actual Length (l)				Ratio of Estimated to Actual Height (h)			
	$\mu_{\ln(l)}$	$\sigma_{\ln(l)}$	$F_{0.05}$	$F_{0.95}$	$\mu_{\ln(h)}$	$\sigma_{\ln(h)}$	$F_{0.05}$	$F_{0.95}$
1	−0.1502	0.4809	0.39	1.90	0.0560	0.6966	0.34	3.33
2	−0.2428	0.5055	0.34	1.80	−0.1308	0.5745	0.34	2.26
3	−0.0188	0.3589	0.54	1.77	0.0028	0.8041	0.27	3.76
4	−0.1249	0.4723	0.41	1.92	0.3340	0.7876	0.38	5.10
5	−0.0200	0.5346	0.41	2.36	0.1192	0.7264	0.34	3.72
6	−0.0038	0.3745	0.54	1.84	−0.1991	0.5785	0.32	2.12
7	0.1585	0.3325	0.68	2.02	0.0668	0.6088	0.39	2.91
8	0.0582	0.3930	0.56	2.02	0.0690	0.6694	0.36	3.22
9	0.0358	0.3657	0.57	1.89	−0.0004	0.7765	0.28	3.59
10	0.1426	0.3272	0.67	1.98	0.2493	0.6994	0.41	4.05
11	0.2572	0.4906	0.58	2.90	0.5474	0.7967	0.47	6.41
12	−0.1766	0.4199	0.42	1.67	−0.1566	0.6870	0.28	2.65
13	0.1574	0.3402	0.67	2.05	0.1882	0.6898	0.39	3.75
14	0.1631	0.2925	0.73	1.90	0.1738	0.6530	0.41	3.48
15	−0.0205	0.3396	0.56	1.71	0.1744	0.5690	0.47	3.04
16	0.0474	0.3804	0.56	1.96	−0.0860	0.8811	0.22	3.91
17	0.2667	0.3680	0.71	2.39	−0.1861	0.6257	0.30	2.32
18	0.0815	0.4493	0.52	2.27	0.2394	0.7547	0.37	4.40
PE	0.0375	0.2652	0.67	1.61	0.0113	0.2127	0.71	1.44

The varying abilities of the NDT technicians to estimate flaw dimensions is apparent in the distributions for flaw length and height that are illustrated in Figure 44. In these figures, the *x*-axis is the ratio of estimated to actual length or height, and the *y*-axis is the probability of exceedance. A value of 1 on the *x*-axis indicates the technician's estimate is equal to the actual length or height. Some technicians tended to under- or overestimate the actual length or height. The technicians who tended to underestimate length or height were those for which the cumulative distribution function intersects the value 1 on the *x*-axis at a value greater than 0.5 on the *y*-axis. For example, technicians 2, 6, 17, and 12 exhibited a tendency to underestimate height, while technicians 11, 4, and 18 exhibited a tendency to overestimate height.

Figure 44. Cumulative distribution functions showing uncertainty in estimates of flaw length and height by technician identifier.



4.7.4 Partial Safety Factors

The lower bound of the confidence interval on the ratio of estimated to actual length or height is of particular interest because it can be used to derive a partial safety factor for FFS analysis. Similar to factors of safety, partial safety factors are designed to provide a target level of reliability without the need for a complete probabilistic analysis. They are called partial safety factors because they apply to an individual input variable, such as flaw length or height, rather than an output variable, such as member capacity. Like safety factors, they depend on both the desired level of reliability and the uncertainty in the input data. The partial safety factor φ is simply the inverse of the lower confidence bound: $\varphi = 1/F_{0.05}$.

Confidence intervals were calculated from the lognormal distributions that were fit to all observations of the ratios of estimated to actual flaw length and height (Figure 36). The lower bound of the 90% confidence interval on the ratio for length was $F_{0.05} = 0.52$ and $\varphi_{\text{length}} = 1.92$. Approximately 95% of flaws will have an actual length that is less than the product of φ_{length} and the estimated length. The lower bound of the 90% confidence interval on the ratio for height was $F_{0.05} = 0.32$ and $\varphi_{\text{height}} = 3.13$. Similarly, 95% of flaws will have an actual height that is less than the product of φ_{height} and the estimated height. A larger value for the partial safety factor for flaw height reflects the fact that flaw height is more difficult to estimate than flaw length. A general rule for applying partial safety factors to FFS analysis is that these results suggest a factor of 2.0 is appropriate for length, and a factor of 3.0 is appropriate for height. These should ensure that a conservative estimate of length and height are being used in FFS analysis about 95% of the time. However, these partial safety factors were based on round-robin experiments conducted in an ideal environment. These may need to be larger for tests conducted under field conditions.

Table 53 shows the partial safety factors that were calculated for different flaw categories and joint types. Partial safety factors for length were consistently around 2.0, with the exception of those for the laminar flaw category and for flaws in straight T-joints. Partial safety factors for height were consistently around 3.0, except for flaws in butt joints and flaws in skewed T-joints. A lower partial safety factor for flaw length or height in any given category suggests those estimates of length and height are less uncertain. However, such conclusions should be supported by an explanation of why there is less uncertainty for a given subgroup of the observations. When the data are subdivided, there are fewer observations to use in estimating

the confidence intervals, and differences among partial safety factors may simply reflect randomness in the data.

Table 53. Partial safety factors for flaw length and height by flaw category and joint type.

Flaw and Specimen Characteristics		Partial Safety Factors	
		Length	Height
Flaw category	Planar	1.92	2.94
	Volumetric	2.04	3.03
	Laminar	1.23	3.03
Joint type	Butt	1.96	2.70
	Corner	2.00	3.03
	Straight T	1.47	3.03
	Skewed T	2.04	7.69
All estimates		1.92	3.13

Partial safety factors were derived for estimates of flaw length and height by technician. These results showed that several technicians produced length or height estimates that had more uncertainty than those of other technicians. Five technicians had partial safety factors that were greater than two for length, and six technicians had partial safety factors that were greater than three for height (Figure 45). Interestingly, only one technician had partial safety factors that exceeded the overall values of two for length and three for height. Table 54 lists the partial safety factors for each technician. The partial safety factors for the technician in the last row of Table 54, for technician PE, were for the Level III technician who performed verification testing of the ERDC specimens. The partial safety factors for this technician were notably lower than for the Level II technicians. In Figure 45, the results for the Level III technician are indicated using a black square, while the results for Level II technicians who participated in the round-robin experiments are indicated using black circles.

Figure 45. Partial safety factors by technician. The *black square* is for the Level III technician, and the *black circles* are for the Level II technicians.

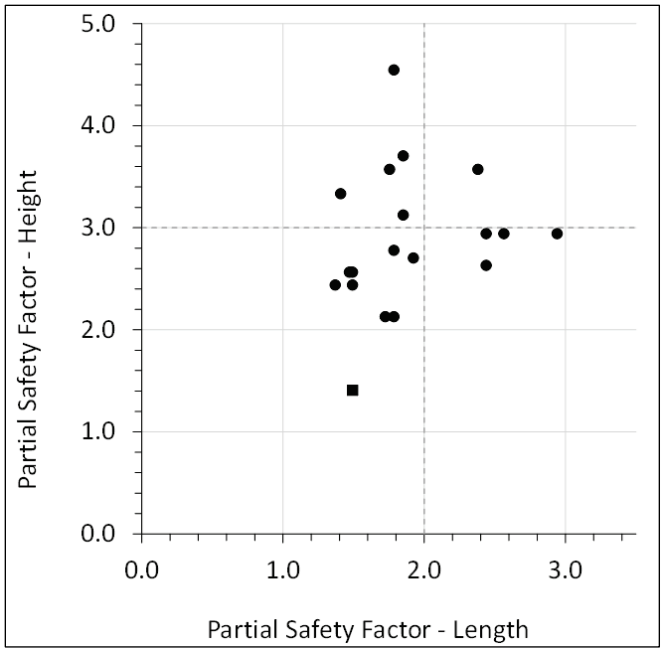


Table 54. Partial safety factors by technician identification (ID).

Technician ID	Partial Safety Factors	
	Length	Height
1	2.56	2.94
2	2.94	2.94
3	1.85	3.70
4	2.44	2.63
5	2.44	2.94
6	1.85	3.13
7	1.47	2.56
8	1.79	2.78
9	1.75	3.57
10	1.49	2.44
11	1.72	2.13
12	2.38	3.57
13	1.49	2.56
14	1.37	2.44
15	1.79	2.13
16	1.79	4.55
17	1.41	3.33
18	1.92	2.70
PE	1.49	1.41

4.8 Flaw Characterization

In some industries, NDT technicians are asked to characterize the discontinuities they detect. Flaw characterization may assist in distinguishing flaws that require repair from those that do not require repair. Technicians were asked to classify flaws into 1 of 10 subcategories: BMCRK, CLCRK, ROCRK, TOCRK, TRCRK, LOP, LOF, POR, SLAG, or LAM. These data were analyzed to compare the accuracy with which technicians could classify flaws using three classification systems. A 10-level classification system was based on the original flaw subcategories. The five different crack subcategories were lumped into a single subcategory (i.e., CRK) to form a six-level classification system: CRK, LOP, LOF, POR, SLAG, and LAM. The six subcategories were lumped according to the three original flaw categories (planar [P], volumetric [V], and laminar [L]) to form a third classification system.

The ability of NDT technicians to accurately classify flaws using each of the three classification systems was assessed using an overall measure of accuracy, ACC, which was calculated as the ratio of the total number of correct classifications across all flaw categories or subcategories to the total number of classification attempts. More detailed classification systems require NDT technicians to make finer distinctions when characterizing flaws. This will be reflected in lower values of ACC for classification systems with more flaw categories or subcategories.

The SEN and PPV were calculated separately for each flaw category and subcategory. SEN was the probability that the flaws belonging to a particular category or subcategory would be correctly classified by the NDT technician. PPV was the probability that a flaw that was classified in a particular category or subcategory belongs within that category or subcategory. SEN was independent of the prevalence of that category or subcategory in the population of flaws. Therefore, estimates of SEN can be used to generalize about the ability of NDT technicians to recognize flaws belonging to a given category or subcategory. However, PPV depends on prevalence. Increasing the prevalence of flaws in a category or subcategory tends to increase the PPV. Therefore, estimates of PPV are specific to a population of flaws.

4.8.1 Three-Level Classification System

Table 55 summarizes the ability of NDT technicians to characterize flaws using the three-level classification system. The row total is the number of attempts to characterize a flaw in each flaw category. The column total is the number of times that the NDT technicians reported a flaw in each flaw category. The number of accurate classifications of flaws in each category is listed in the cells on the diagonal, and the total number of observations is listed in the lower right-hand cell. This table shows that there were 51 attempts to classify laminar (L) flaw and that NDT technicians accurately reported the characterization of laminar flaws in 41 of those attempts. Laminar flaws were misclassified as planar (P) flaws in 9 instances and as volumetric (V) flaws in 1 instance. Similarly, NDT technicians accurately reported the characterization of planar flaws in 618 out of 704 instances and volumetric flaws (V) in 125 of 226 instances. Overall, flaw classification accuracy was 79.9% for the three-level classification system.

Table 55. Reported and actual flaw characterization for the classification system with three flaw categories.

Actual Flaw Category	Reported Flaw Category				
	L	P	V	Row Total	SEN
L	41	9	1	51	0.804
P	2	618	84	704	0.878
V	1	100	125	226	0.553
Column total	44	727	210	981	—
PPV	0.932	0.850	0.595	—	ACC: 0.799

SEN is the probability that a flaw of a given category will be correctly classified. SEN is listed on the right-hand side of the table. In percentage terms, technicians accurately classified laminar flaws 80.4% of the time (41 of 51 attempts), planar flaws 87.8% of the time (618 of 704 attempts) and volumetric flaws 55.3% of the time (125 of 226 attempts). The figure shows that volumetric flaws were often misclassified as planar flaws. This tendency can probably be attributed to NDT technicians having been trained to classify flaws as planar when they are at all uncertain about the true character of that flaw. SEN does not depend on the prevalence of each flaw type in a population.

PPV is the probability that the characterization of a flaw reported by a technician is correct. These values are listed for each reported flaw

category in the last row of the table. The probability that a flaw was a laminar flaw given that the technician reported that it was a laminar flaw was 0.932. Similarly, the PPV for planar flaws was 0.85, and the PPV for volumetric flaws was 0.595. In contrast to SEN, the PPV for a given category depends on the prevalence of that flaw category in a population. Therefore, estimates of PPV should not be extrapolated to a different population.

4.8.2 Six-Level Classification System

The six-level classification system included three categories for planar flaws (i.e., CRK, LOF, and LOP), two categories for volumetric flaws (i.e., SLAG and POR), and one category for laminar flaws (i.e., LAM). Table 56 summarizes the ability of technicians to characterize detected flaws using the six-category system. For the classification system as a whole, NDT technicians accurately classified 59.2% of detected flaws (ACC = 0.592).

Table 56. Reported and actual flaw characterization for the classification system with six flaw categories.

Actual Flaw Category and Subcategory		Reported Flaw Category and Subcategory							
		L	P			V		Row Total	SEN
		LAM	CRK	LOF	LOP	POR	SLAG		
L	LAM	41	5	3	1	0	1	51	0.804
P	CRK	2	284	42	23	6	26	383	0.742
	LOF	0	28	76	19	9	19	151	0.503
	LOP	0	38	19	72	3	17	149	0.483
V	POR	0	7	4	0	46	19	76	0.605
	SLAG	1	45	32	9	10	44	141	0.312
Column total		44	407	176	124	74	126	951	—
PPV		0.932	0.698	0.432	0.581	0.622	0.349	—	ACC: 0.592

SEN is listed for each flaw category in the right-most column of the table, and PPV is listed for each reported flaw category in the bottom row of the table. The results for laminar flaws are the same as in the three-level classification system because the laminar flaw category contains a single subcategory. For flaws characterized as CRK, SEN = 0.742. For flaws reported as CRK, PPV was 0.698. Both SEN and PPV were notably lower for LOF and LOP. For flaws characterized as LOF, SEN = 0.503, and for flaws reported to be LOF, PPV = 0.432. In other words, flaws that were reported to be LOF were more likely to be something else. For flaws characterized as LOP, SEN = 0.483 and PPV = 0.581. For POR and SLAG, SEN was 0.605

and 0.312, respectively, and PPV was 0.622 and 0.349, respectively. There was a pronounced tendency to report SLAG as either CRK or LOF. One possible explanation for this is that SLAG is often associated with LOF; however, there was a greater tendency to classify SLAG as CRK than LOF. Note that the number of total observations in Table 56 (951 observations) differs from that in Table 55 (981 observations) because there were 30 instances in which the NDT technician declined to classify detected flaws using the more refined classification system.

4.8.3 Ten-Level Classification System

The 10-level classification system was the most difficult of the three systems because NDT technicians were required to distinguish between different types of cracks. Table 57 summarizes the results. In the table, the five different types of cracks are referenced using only the first two letters of their abbreviation under the column subheading CRK. The overall accuracy for the 10-level classification system was 0.507.

Table 57. Reported and actual flaw characterization for the classification system with 10 levels.

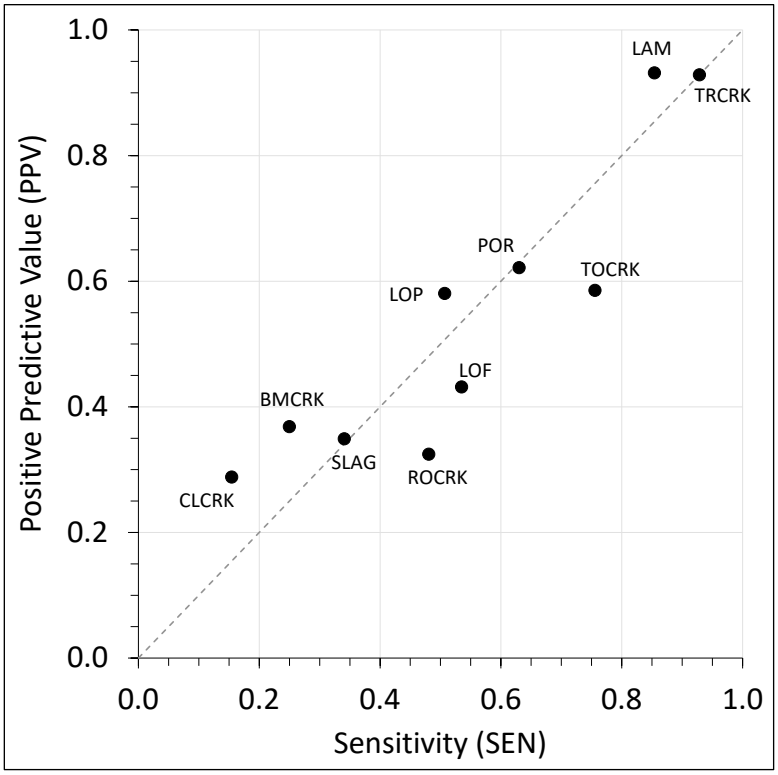
Actual Flaw Category and Subcategory			Reported Flaw Category and Subcategory										Row Total	SEN
			L	P							V			
				CRK										
			LAM	BM	CL	RO	TO	TR	LOF	LOP	POR	SLAG		
L	LAM		41	1	0	0	0	1	3	1	0	1	48	0.854
P	CRK	BM	1	7	1	2	14	0	1	0	1	1	28	0.250
		CL	0	2	17	10	16	2	25	15	5	18	110	0.155
		RO	0	0	2	25	3	0	9	8	0	5	52	0.481
		TO	0	8	2	2	65	0	7	0	0	2	86	0.756
		TR	1	0	1	0	1	39	0	0	0	0	42	0.929
	LOF		0	0	3	8	8	0	76	19	9	19	142	0.535
	LOP		0	0	17	12	2	0	19	72	3	17	142	0.507
V	POR		0	1	2	1	0	0	4	0	46	19	73	0.630
	SLAG		1	0	14	17	2	0	32	9	10	44	129	0.341
Column total			44	19	59	77	111	42	176	124	74	126	852	
PPV			0.932	0.368	0.288	0.325	0.586	0.929	0.432	0.581	0.622	0.349		ACC: 0.507

For subcategories other than CRK, SEN and PPV values were similar to those in the three- and six-level classification systems. With respect to CRK, NDT technicians had the most difficulty recognizing BMCRKs and CLCRKs. Technicians characterized CLCRKs accurately in 17 out of 110

attempts, or 15.5% of the time. CLCRKs were often mistaken for a ROCRK, TOCRK, LOF, LOP, or SLAG. BMCRKs were accurately characterized in 7 out of 28 attempts, or 25% of the time. BMCRKs were often confused with TOCRKs. In contrast, technicians accurately characterized TRCRKs in 39 out of 42 attempts (92.9% of the time) and TOCRKs in 65 out of 86 attempts (75.6% of the time). The PPV for different types of cracks was notably lower than for other flaw categories, except for TRCRK. The PPV for TRCRK was 0.929.

Figure 46 is a plot of SEN and PPV. It shows how much more difficult it is to classify some types of flaws than others. It also shows that SEN and PPV appear to be correlated, which suggests that Type I errors (i.e., FP) and Type II errors (i.e., FN) are occurring in roughly equal proportions.

Figure 46. Sensitivity (SEN) and positive predictive value (PPV) for each flaw category in the 10-level classification system.



4.8.4 Variability in Accuracy and Reliability among NDT Technicians

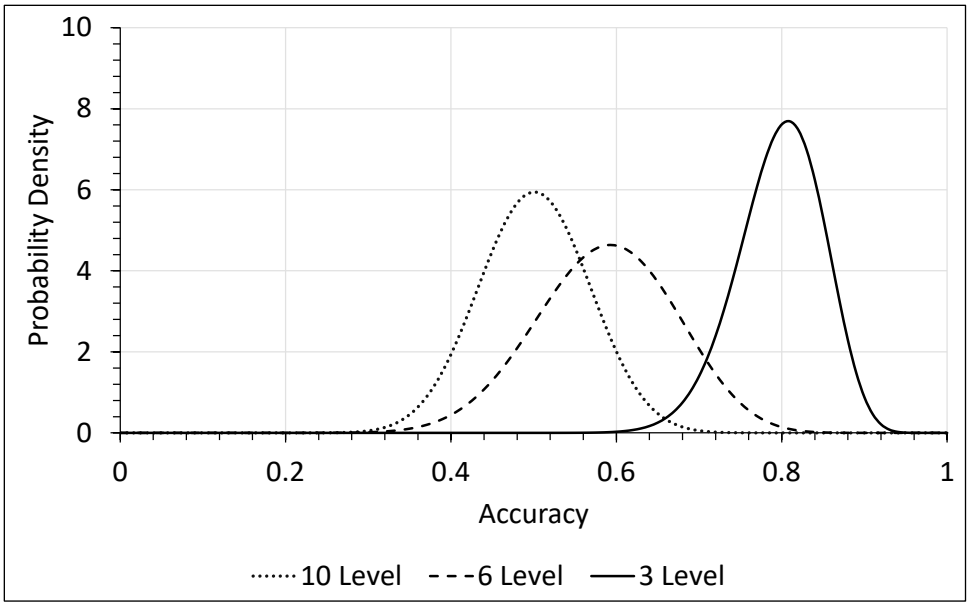
The accuracy with which NDT technicians classified flaws varied with technician and classification system (Table 58). For the classification system consisting of three levels, technician accuracy ranged from 0.72 to 0.90. For the classification system consisting of six levels, technician

accuracy ranged from 0.43 to 0.75. For the classification system consisting of 10 levels, technician accuracy ranged from 0.40 to 0.65. Table 58 lists the number of flaws classified by each technician. This number varied by technician because each technician could only characterize a flaw after it had been detected, and each technician detected a different number of flaws. Beta density functions in Figure 47 illustrate the effect of the classification system on the accuracy of flaw characterization. The expected accuracy was lowest for the 10-level classification system.

Table 58. Fraction of flaws detected and characterized accurately using the 10-, 6-, and 3-level classification systems.

	ACC			Number of Flaws Classified		
Technician ID	10 Levels	6 Levels	3 Levels	10 Levels	6 Levels	3 Levels
1	0.40	0.43	0.72	40	44	46
2	0.40	0.46	0.75	43	48	48
3	0.50	0.55	0.75	42	53	53
4	0.40	0.47	0.73	47	53	55
5	0.47	0.55	0.78	36	51	51
6	0.40	0.50	0.86	42	44	56
7	0.54	0.63	0.73	41	56	56
8	0.52	0.66	0.86	33	47	49
9	0.60	0.70	0.81	55	56	57
10	0.52	0.63	0.77	48	57	57
11	0.52	0.56	0.75	46	52	56
12	0.48	0.53	0.80	33	40	40
13	0.49	0.62	0.80	55	55	59
14	0.54	0.63	0.82	56	57	57
15	0.53	0.66	0.82	57	58	60
16	0.65	0.75	0.90	57	59	60
17	0.54	0.61	0.87	61	61	61
18	0.52	0.63	0.83	60	60	60
Overall	0.50	0.59	0.80	852	951	981

Figure 47. Variability in flaw classification accuracy across nondestructive testing (NDT) technicians for each of the three classification systems.



5 Application of Partial Safety Factors to Fitness-for-Service (FFS) Examples

Guidance for USACE engineers performing structural evaluations of in-service HSS is provided in EM 1110-2-6054 (USACE 2001), *Inspection, Evaluation, and Repair of Hydraulic Steel Structures*, and in Dexter et al.'s (2007), *Fitness-for-Purpose Evaluation of Hydraulic Steel Structures*. Dexter et al. (2007) provided a general procedure for FFS evaluations of HSS based on API 579, *Fitness-for-Service*, and BS-7910:2013+A1:2015, *Guide to Methods for Assessing the Acceptability of Flaws in Metallic Structures*. EM 1110-2-6054 (USACE 2001) is currently being revised, but those revisions were not available at the time of writing. The examples presented in this chapter use the BS-7910:2013+A1:2015 (BSI 2015) approach to FFS evaluations because that is the approach adopted in EM 1110-2-6054. These examples focus on the application and effects of partial safety factors applied to the reported defect size to account for inaccuracies in flaw sizing using PAUT. A complete evaluation of in-service HSS includes the assessment of both long- and short-term effects of all reported damage under all expected loading conditions in accordance with applicable EMs and industry standards.

5.1 Failure Assessment Diagram (FAD)

An in-depth discussion of the theory and background of the failure assessment diagram (FAD) approach is provided in Dexter et al. (2007) and in BS-7910:2013+A1:2015 (BSI 2015). In short, the FAD presents a failure envelope that separates acceptable defects from unacceptable defects considering failure due to brittle fracture and plastic collapse. The vertical axis represents the potential for brittle fracture, the horizontal axis represents the potential for plastic collapse, and the failure envelope represents the interaction between the two failure modes. The assessment considers member stress; crack type, size, and location; and material properties. BS-7910:2013+A1:2015 includes three assessment levels, designated Option 1, Option 2, and Option 3. Option 1, the least complex approach, does not require detailed stress–strain data for the metal. Option 2 requires stress–strain data to define the failure envelope. Finally, Option 3, the most complex approach, requires both detailed stress–strain data and an elastic and elastic–plastic analysis of the flawed structure. Option 3 is generally considered only if Options 1 or 2 do not show satisfactory results. Figures 48 and 49, respectively, show an example FAD from an Option 1 and an

Option 2 assessment. This chapter is intended for engineers that are familiar with the Option 1 assessment procedure.

Figure 48. Schematic representation of an Option 1 failure assessment diagram (FAD).

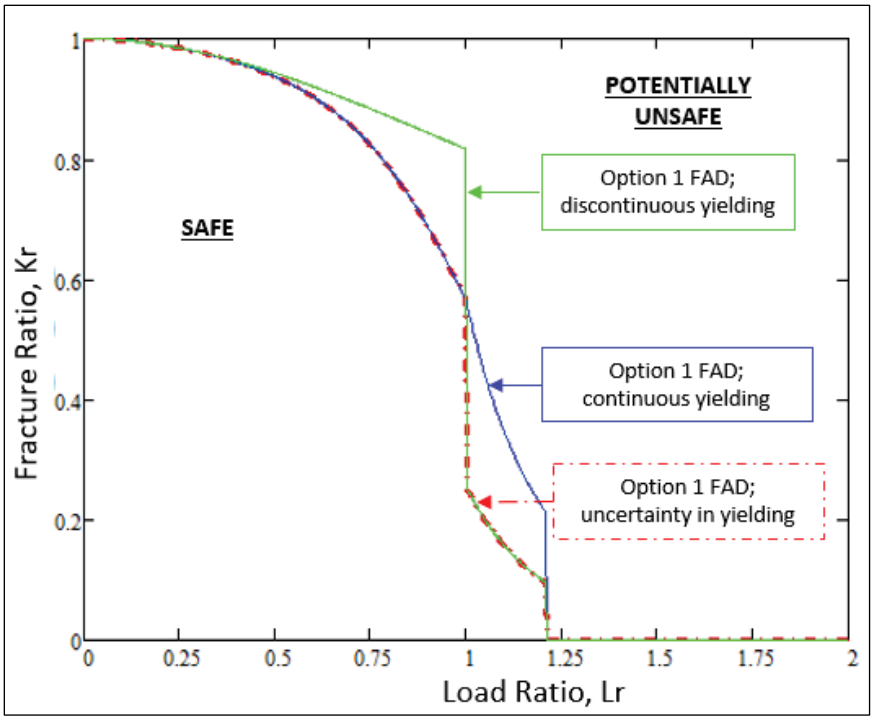
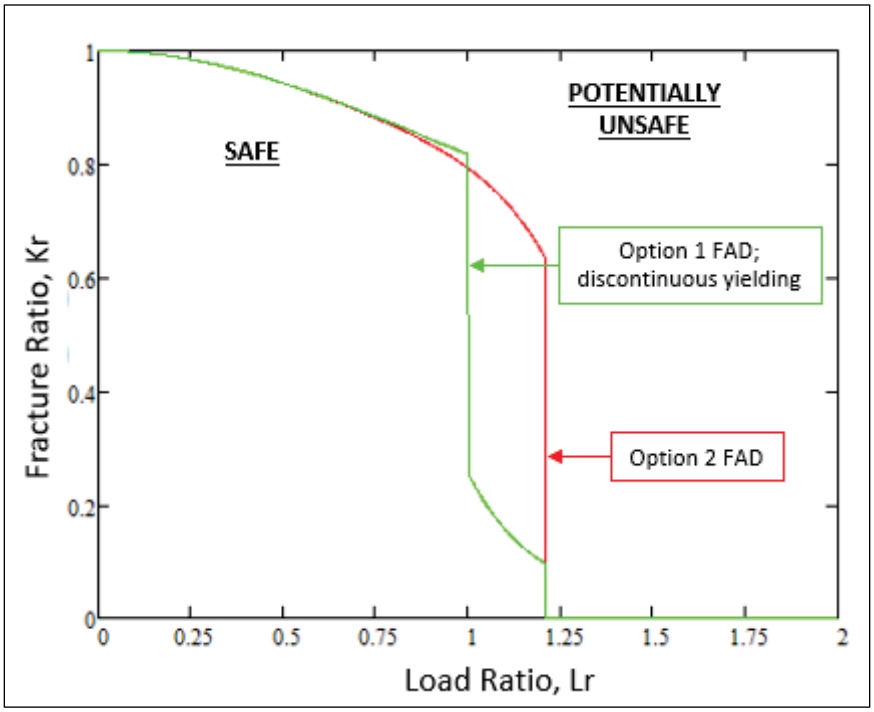


Figure 49. Schematic representation of an Option 2 FAD.



5.2 Partial Safety Factors

No inherent safety factors are included in the development of the FAD. Appropriate safety factors must be applied by the engineer to obtain the required level of reliability. BS-7910:2013+A1:2015 (BSI 2015), Appendix K, provides two levels of reliability analysis. Level I is a semiprobabilistic approach based on the application of partial safety factors, and Level II is a probabilistic approach based on the first order reliability methods. BS-7910:2013+A1:2015 provides recommendations for partial safety factors for stress, flaw size, and material toughness and yield strength for target levels of reliability. For flaw size, the partial safety factors were derived based on the inspection capabilities outlined in Tables T.1 through T.3 and the studies summarized in ENA TS 98/10 (ENA [1983] in BSI [2015]) and Offshore Technology Report 2000/020 (Burdekin and Hamour 2003). Partial safety factors for defect size in BS-7910:2013+A1:2015 (BSI 2015) range from 1.0 to 2.5, depending on the target level of reliability and the coefficient of variance. Different partial safety factors for flaw height and flaw length are not provided, and the partial safety factors are intended to apply to all methods of NDT. Based on the results from the round-robin inspection presented in Chapter 4, flaw detection and sizing from NDT of in-service HSS is not within the accuracy ranges used to calculate the partial safety factor for flaw size in BS-7910:2013+A1:2015.

This example considers the application of a partial safety factor to the reported flaw size based on the findings from the round-robin inspections. As recommended in Section 4.7.4 of this report, a partial safety factor of 3.0 will be applied to the reported height, and a partial safety factor of 2.0 will be applied to the reported length. These factors of safety were not calibrated to provide a target level of reliability. To achieve this, the safety factors for load and resistance effects must be calibrated simultaneously for HSS. The updated version of EM 1110-2-6054 is expected to include additional discussion on implementation of safety factors for FFS evaluations of HSS. When the FFS evaluation indicates marginal acceptance or rejection, a sensitivity analysis should be performed to assess the effects of changes in the input data. Overly conservative assumptions on the probability of failure will lead to a greater likelihood of assessment failures.

5.3 Fracture Evaluation Example 1: Flat Plate with Surface Crack

The influence of partial safety factors for a fracture evaluation of a flat plat with a surface crack will be investigated using the example proposed for

the revision to EM 1110-2-6054. This example was developed by Michael Baker International and Southwest Research Institute to demonstrate the use of NASGRO software (SwRI and NASA 2022a) for FFS evaluations. The example selected for EM 1110-2-6054 uses the FITNET FAD implementation. This approach is similar, but not identical, to the BS-7910:2013+A1:2015 (BSI 2015) implementation. A comparison of the two approaches is included in this example. Complete calculations are provided in Appendix F.

5.3.1 Example Problem Description

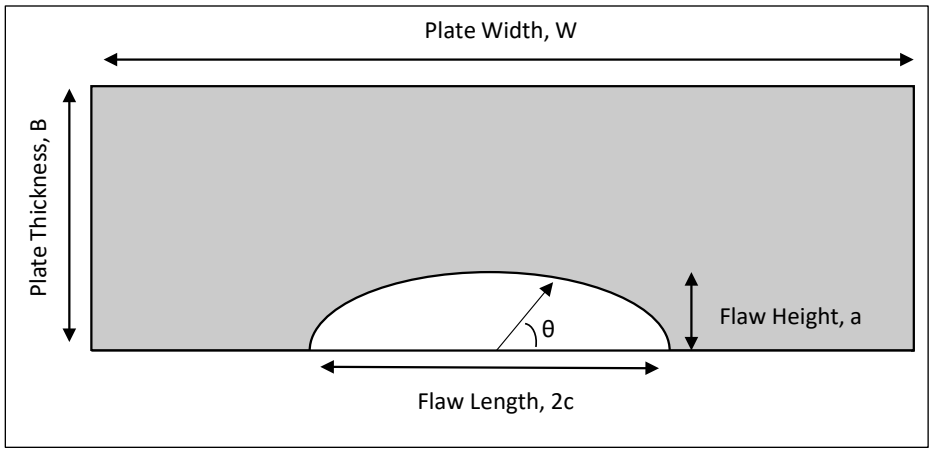
Perform a fracture mechanics evaluation using the BSI (2015) BS 7910:2013+A1:2015 Option 1 (FITNET Option 1) for this problem (Figure 50):

- Material properties
    - Material: Carbon structural steel with matching filler material
    - Minimum yield strength (σ_y): 44 ksi
    - Minimum tensile strength (σ_u): 78 ksi
    - Lower bound fracture toughness (K_{mat}): 60 ksi√in.
    - Plate thickness (B): 0.75 in.
    - Plate width (W): 12 in.
- Loading
    - Applied membrane stress (P_m): 15 ksi
    - Applied bending stress (P_b): 0 ksi
    - Residual membrane stress (Q_m): 44 ksi (uniform through thickness)
    - Residual bending stress (Q_b): 0 ksi
- Discontinuities
    - Semi-elliptical surface crack (centered along the width of plate)
        - Crack height (a)*: 0.2 in.
        - Crack length ($2c$)*: 0.6666 in.
        - Crack height (a): 0.2 in.
        - Crack length ($2c$): 0.46 in.
        - Crack height (a): 0.06 in.
        - Crack length ($2c$): 0.6666 in.

*Discontinuity used for example calculations.

Note: All inputs are theoretical and should not be used in the evaluation of in-service HSS. Material properties and member stresses for FFS evaluations of HSS should be determined in accordance with the recommendations of EM 1110-2-6054 (USACE 2001) and industry standards.

Figure 50. Plate and flaw geometry for a surface flaw. (Image adapted from BS-7910:2013+A1:2015 [BSI 2015], Figure M.3.)



5.3.2 Calculate Load Ratio, L_r

The load ratio is calculated as the ratio of the reference stress, σ_{ref}, to the material yield stress, σ_y, where the reference stress represents the loading condition being assessed, and the yield stress represents the limit load of the structure. The load ratio describes the structure's proximity to plastic collapse. Unique solutions for reference stress have been developed for a variety of flaw types. For a flat plate containing surface flaws, the reference stress can be calculated in accordance with equations P.9 and P.10 from BS-7910:2013+A1:2015 (BSI 2015). For this example, normal bending restraint was assumed, and the reference stress was determined to be 16.3 ksi. The load ratio was then determined to be 0.371.

$$L_r = \frac{\sigma_{ref}}{\sigma_y} = \frac{16.341 \text{ ksi}}{44 \text{ ksi}} = 0.371.$$

NASGRO (SwRI and NASA 2022b) defines the load ratio as the ratio of the total applied load to the plastic limit load, and a value of 0.35 was calculated for the load ratio.

5.3.3 Calculate Fracture Ratio, K_r

The fracture ratio is defined as the ratio of the stress intensity factor, K_I, to the material fracture toughness, K_{mat}. A plasticity correction factor, ρ, is included to account for the interaction of primary and secondary stress. The fracture ratio represents the structure's proximity to brittle fracture. The stress intensity factor is a function of primary and secondary stresses and flaw size. The primary and secondary stresses are multiplied by magnification factors to account for stress concentrations due to misalignment, flaw geometry and location, welded joint geometry, and so on.

Unique stress intensity solutions have been developed for a range of flaw types. For a flat plate with a surface or embedded flaws, the stress intensity factors for both membrane (i.e., axial) loading and bending loading can be calculated in accordance with BS-7910:2013+A1:2015 (BSI 2015), Appendix M.4. For this flaw, the maximum stress intensity factor, K_I = 40.9 ksi√in., occurs at a point along the crack tip perpendicular to the member surface ($\theta = 90°$). The simplified procedure from Appendix R.2 was used to calculate the plasticity correction factor. The fracture ratio was determined to be 0.779.

$$K_r = \frac{K_I^P + K_I^S}{K_{mat}} + \rho = \frac{30.5 \text{ ksi}\sqrt{\text{in.}} + 10.4 \text{ ksi}\sqrt{\text{in.}}}{60 \text{ ksi}\sqrt{\text{in.}}} + 0.098 = 0.779.$$

NASGRO (SwRI and NASA 2022a) uses the FITNET implementation for the plasticity correction factor, which requires interpolation from tables in lieu of the equations provided in BS-7910:2013+A1:2015 (BSI 2015). Using the FITNET tables, the plasticity correction factor, ρ, is calculated to be 0.11, and the fracture ratio is calculated to be 0.787. Tables for interpolation are provided in Appendix X of the NASGRO user's manual (SwRI and NASA 2022b).

5.3.4 Develop FAD

For this example, the FAD was developed for Option 1 with continuous yielding. The equations to define the curve are provided in Section 7.3 of BS-7910:2013+A1:2015 (BSI 2015). The curve is defined by $K_r = f(L_r)$, with a cutoff value of $L_{r,max} = (\sigma_y + \sigma_u)/2\sigma_y$. The maximum value for the fracture ratio is 1.0. The structural steel used in HSS typically exhibits discontinuous yielding; however, NASGRO does not have the capability to evaluate a discontinuously yielding material at this time (SwRI and NASA 2022b).

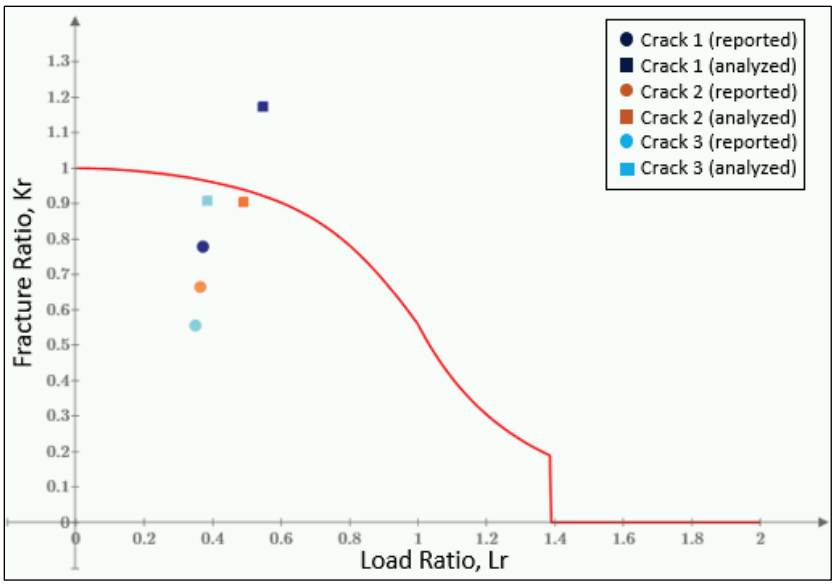
In this example, three cracks are evaluated, with and without partial safety factors applied to the reported defect size. In Figure 51, assessment points based on the reported flaw size (without partial safety factor) are shown in circles, and the assessment points for the analyzed flaw size (with partial safety factor) are shown in squares. Crack 1 is in dark blue, Crack 2 is in orange, and Crack 3 is in light blue. An examination of the FAD for these three cracks indicates that for Crack 1, the assessment point is outside the safe region when partial factors of safety are applied. For Cracks 2 and 3, the assessment point is within the safe region when the partial safety factors are applied, although the points are located on the margins of the envelope. It may be possible to demonstrate the acceptability of Crack 1 by applying a higher assessment option. Otherwise, Crack 1 should be repaired, or the loading on the structure should be reduced to an acceptable level.

In some cases, application of the partial factors of safety may change the flaw type (e.g., from surface or embedded to through thickness), or it may cause the location of the maximum stress intensity factor along the crack tip to shift. This shift is not reflected in Table 59 or Figure 51.

Table 59. Summary of results for three surface cracks in a carbon steel plate.

Crack No.	Crack Height, a (in.)	Half of Crack Length, c (in.)	Load Ratio, L_r	Fracture Ratio, K_r
1 (reported)	0.2	0.3333	0.371	0.779
1 (analyzed)	0.6	0.6666	0.547	1.174
2 (reported)	0.2	0.23	0.364	0.665
2 (analyzed)	0.6	0.46	0.49	0.905
3 (reported)	0.06	0.3333	0.35	0.555
3 (analyzed)	0.18	0.6666	0.384	0.909

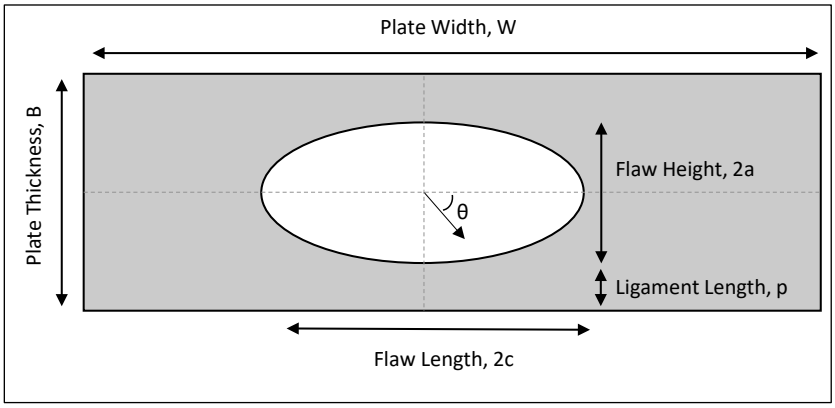
Figure 51. FAD for three surface cracks in a carbon steel plate. Assessment points based on the reported flaw size (without partial safety factors) are *circles*, and assessment points for the analyzed flaw size (with partial safety factors) are *squares*.



5.4 Fracture Evaluation Example 2: Flat Plate with Embedded Crack

The influence of partial safety factors for a fracture evaluation of a flat plate with an embedded crack will be investigated in this example (Figure 52). This example is not included in the revisions to EM 1110-2-6054, so NASGRO analysis software (SwRI and NASA 2022a) was used to validate the results. In NASGRO, the FITNET FAD implementation was used. This approach is similar, but not identical, to the BS-7910:2013+A1:2015 (BSI 2015) implementation. A comparison of the two approaches is included in this example. Appendix F contains the complete calculations, including the output from NASGRO.

Figure 52. Plate and flaw geometry for an embedded flaw. (Image adapted from BS-7910:2013+A1:2015 [BSI 2015], Figure M.8).



5.4.1 Example Problem Description

Perform a fracture mechanics evaluation using the BS-7910:2013+A1:2015 (BSI 2015) Option 1 (FITNET Option 1) for this example problem:

- Material properties
    - Material: Carbon structural steel with matching filler metal
    - Minimum yield strength (σ_y): 36 ksi
    - Minimum tensile strength (σ_u): 58 ksi
    - Lower bound fracture toughness (K_{mat}): 60 ksi√in.
    - Plate thickness (B): 0.75 in.
    - Plate width (W): 12 in.

- Loading
    - Applied membrane stress (P_m): 15 ksi
    - Applied bending stress (P_b): 5 ksi
    - Residual membrane stress (Q_m): 36 ksi (uniform through thickness)
    - Residual bending stress (Q_b): 0 ksi

- Discontinuities
    - Semielliptical embedded crack (centered along the width and thickness of the plate)
        - Crack height ($2a$)*: 0.2 in.
        - Crack length ($2c$)*: 0.6666 in.
        - Ligament length (p)*: 0.275 in.
        - Crack height ($2a$): 0.2 in.
        - Crack length ($2c$): 0.4 in.
        - Ligament length (p): 0.275 in.
        - Crack height ($2a$): 0.14 in.
        - Crack length ($2c$): 0.6666 in.
        - Ligament length (p): 0.305 in.

*Discontinuity used for example calculations.

Note: All inputs are theoretical and should not be used in the evaluation of in-service HSS. Material properties and member stresses for FFS

evaluations of HSS should be determined in accordance with the recommendations of EM 1110-2-6054 (USACE 2001) and industry standards.

5.4.2 Calculate Load Ratio, L_r

The load ratio is calculated as the ratio of the reference stress, σ_{ref}, to the material yield stress, σ_y, where the reference stress represents the loading condition being assessed, and the yield stress represents the limit load of the structure. The load ratio describes the structure's proximity to plastic collapse. Unique solutions for reference stress have been developed for a variety of flaw types. For a flat plate containing embedded flaws, the reference stress can be calculated in accordance with equation P.11 from BS-7910:2013+A1:2015 (BSI 2015). For this example, normal bending restraint was assumed, and the reference stress was determined to be 18.6 ksi. The load ratio was then determined to be 0.516.

$$L_r = \frac{\sigma_{\text{ref}}}{\sigma_y} = \frac{18.588 \text{ ksi}}{36 \text{ ksi}} = 0.516.$$

NASGRO (SwRI and NASA 2022b) defines the load ratio as the ratio of the total applied load to the plastic limit load and a value of 0.47 was calculated for the load ratio.

5.4.3 Calculate Fracture Ratio, K_r

The fracture ratio is defined as the ratio of the stress intensity factor, K_I, to the material fracture toughness, K_{mat}. A plasticity correction factor, ρ, is included to account for the interaction of primary and secondary stress. The fracture ratio represents the structure's proximity to brittle fracture. The stress intensity factor is a function of primary and secondary stresses and flaw size. The primary and secondary stresses are multiplied by magnification factors to account for stress concentrations due to misalignment, flaw geometry and location, welded joint geometry, and so on.

Unique stress intensity solutions have been developed for a range of flaw types. For a flat plate with a surface or embedded flaws, the stress intensity factors for both membrane (i.e., axial) loading and bending loading can be calculated in accordance with BS-7910:2013+A1:2015 (BSI 2015), Appendix M.4. For this flaw, the maximum stress intensity factor, K_I = 26.8 ksi√in., occurs at a point along the crack tip perpendicular to the major axis of the flaw ($\theta = 90°$). The simplified procedure from Appendix R.2

was used to calculate the plasticity correction factor. The fracture ratio was determined to be 0.551.

$$K_{\mathrm{r}} = \frac{K_I^P + K_I^S}{K_{\mathrm{mat}}} + \rho = \frac{18.7\ \mathrm{ksi}\sqrt{\mathrm{in.}} + 8.1\ \mathrm{ksi}\sqrt{\mathrm{in.}}}{60\ \mathrm{ksi}\sqrt{\mathrm{in.}}} + 0.103 = 0.551.$$

Note that NASGRO (SwRI and NASA 2022a) uses the FITNET implementation for the plasticity correction factor, which requires interpolation from tables in lieu of the equations provided in BS-7910:2013+A1:2015 (BSI 2015). Using the FITNET tables, the plasticity correction factor was calculated to be 0.12, and the fracture ratio was calculated to be 0.569. Tables for interpolation are provided in Appendix X of the NASGRO user's manual (SwRI and NASA 2022b).

5.4.4 Develop FAD

For this example, the FAD was developed for Option 1 with discontinuous yielding. The equations to define the curve are provided in Section 7.3 of BS-7910:2013+A1:2015 (BSI 2015). The curve is defined by $K_{\mathrm{r}} = f(L_{\mathrm{r}})$, with a cutoff value of $L_{\mathrm{r,max}} = (\sigma_{\mathrm{y}} + \sigma_{\mathrm{u}})/2\sigma_{\mathrm{y}}$. The maximum value for the fracture ratio is 1.0. The FAD for continuous yielding was also plotted.

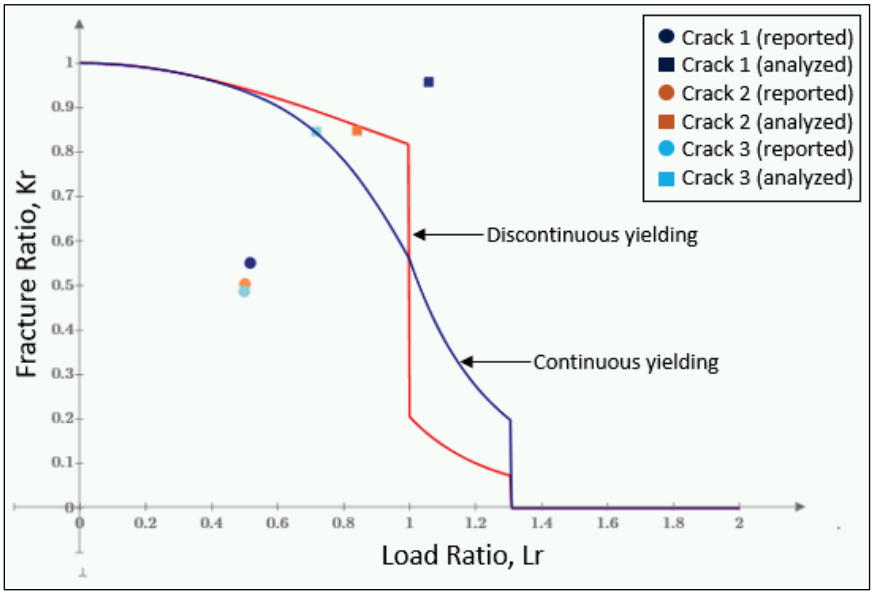
In this example, three cracks are evaluated, with and without partial safety factors applied to the reported crack size. In Figure 53, assessment points based on the reported flaw size (without partial safety factor) are shown in circles, and the assessment points for the analyzed flaw size (with partial safety factor) are shown in squares. Crack 1 is in dark blue, Crack 2 is in orange, and Crack 3 is in light blue. The FAD assuming discontinuous yielding is shown in red, and the FAD for continuous yielding is shown in blue. Most structural steels exhibit a yield discontinuity or plateau where necking occurs prior to strain hardening. Therefore, the FAD for discontinuous yielding is used for evaluation. Examination of the FAD for these three cracks indicates that for Crack 1, the assessment point is outside the safe region when partial factors of safety are applied. For Cracks 2 and 3, the assessment point is within the safe region when the partial safety factors are applied, although the points are located on the margins of the envelope. It may be possible to demonstrate the acceptability of Crack 1 by applying a higher assessment option. Otherwise, Crack 1 should be repaired, or the loading on the structure should be reduced to an acceptable level.

In some cases, application of the partial factors of safety may change the flaw type (e.g., from surface or embedded to through thickness), or it may cause the location of the maximum stress intensity factor along the crack tip to shift. This shift is not reflected in Table 60 or Figure 53.

Table 60. Summary of results for three embedded cracks in a carbon steel plate.

Crack No.	Half of Crack Height, *a* (in.)	Half of Crack Length, *c* (in.)	Ligament Length, *p* (in.)	Load Ratio, L_r	Fracture Ratio, K_r
1 (reported)	0.1	0.3333	0.305	0.551	0.516
1 (analyzed)	0.3	0.6666	0.075	0.958	1.058
2 (reported)	0.1	0.2	0.275	0.504	0.5
2 (analyzed)	0.3	0.4	0.075	0.848	0.84
3 (reported)	0.07	0.3333	0.305	0.487	0.498
3 (analyzed)	0.21	0.6666	0.165	0.846	0.716

Figure 53. FAD for three embedded cracks in a carbon steel plate. Assessment points based on the reported flaw size (without partial safety factors) are *circles*, and the assessment points for the analyzed flaw size (with partial safety factors) are *squares*.



6 Discussion and Recommendations

This project was motivated, in part, by ongoing efforts to update EM-1110-2-6054 (USACE 2001). That document, along with ER 1110-2-100 (USACE 1995) and ER 1110-2-8157 (USACE 2009), provides guidance on the inspection, evaluation, and repair of HSS. Updates are needed to address advancements in UT techniques and FFS analysis. A review of UT techniques concluded that together, PAUT, TFM/FMC, and TOFD provide the flexibility required to conduct testing effectively in the wide variety of joint geometries encountered in existing HSS. A review of FFS methodologies concluded that BS 7910:2013+A1:2015 (BSI 2015) is the most generally applicable method of FFS for HSS. During 2019, BSI released an updated version of BS 7910 (BSI 2019). This version incorporated much of the existing material and, because one of the objectives was to maintain continuity with previous versions of the methodology, the methods described in BS 7910:2013+A1:2015 (BSI 2015) remain valid.

BS 7910:2013+A1:2015 (BSI 2015) is most effective when used in conjunction with information on the size and character of subject flaws and actual material properties. The research described in this technical report is aimed at understanding the ability to detect, size, and characterize flaws using PAUT. Logistic regression models were fit to the results of a round-robin experiment to estimate POD and identify factors that influence the ability to detect flaws. Proportional errors in estimates of flaw length and height were modeled using lognormal distributions to characterize uncertainty in those estimates. Partial safety factors were derived to help account for those uncertainties in FFS analysis. Two examples of FFS analysis were provided to demonstrate how those partial safety factors can be used. The ability of NDT technicians to characterize flaws using PAUT was also analyzed. This chapter summarizes what was learned through this research project and concludes with several recommendations that should be considered for incorporation into proposed revisions of EM 1110-2-6054.

6.1 NDT Techniques and Procedures

NDT is a collection of techniques for detecting, sizing, and characterizing welding defects and discontinuities without causing damage to the structure. Various NDT techniques were considered for use on HSS. UT is one form of NDT that has been widely used in evaluating heavy civil structures

such as pipelines, bridges, and HSS. This study identified three UT techniques for use on HSS: PAUT, TFM/FMC, and TOFD. Discussions with industry experts indicated that these techniques are effective and widely used. One advantage of these techniques is that they generate a permanent record of every scan, which can be independently reviewed by other technicians. Modern UT instruments are capable of performing all three techniques, and because these techniques have complimentary advantages and disadvantages, this improves reliability. The research team concluded that these three techniques would provide the availability, convenience, flexibility, and reliability for work on HSS.

ASNT Level III certified NDT technicians developed NDT testing procedures and scan plans for each of these NDT techniques. The procedures and scan plans were validated and improved during a week of testing. While some in the industry have promoted the idea that PAUT line scans are sufficient to detect and size flaws, our results demonstrated that, particularly with respect to the weld specimens with more complicated geometries, line scans are not adequate. Other authors have reached similar conclusions (Connor et al. 2019). The quality of inspection results improved with the addition of several elements to the procedures. These elements included manual rastering, line scans from multiple index offsets, scanning from all faces, and using smaller probes. These results also demonstrated that skills vary widely among NDT technicians and that expensive instruments cannot compensate for technicians with poor skills. For example, the ASNT Level III technician who conducted verification testing of the specimens used PE to detect and size flaws with less error than the technicians who participated in round-robin experiments using PAUT. PE requires a much cheaper instrument than does PAUT.

6.2 Finding NDT Technicians

The research team worked with Michael Baker International to identify NDT technicians and recruit them to participate in round-robin testing. Identifying technicians who were qualified to implement at least two of the three selected UT techniques and willing to participate in a blind testing research project was much more difficult than anticipated. Although PAUT was added to the AWS codes that govern NDT in the structural industry in 2015, few technicians working in the structural industry had experience using these techniques. In addition, NDT technicians in the structural industry do not practice height sizing or flaw characterization because AWS

does not require it. The AWS rejection criteria are amplitude based, with additional criteria based on length but not height. The research team also found that state-level DOTs have a wide variety of practices for qualifying NDT technicians (e.g., California DOT [2021] and New York Department of Labor [2021]). Differences in training and testing practices across the states have produced disparities in skill levels among NDT technicians in the structural industry.

PAUT is widely used for pipeline inspections, and the American Society of Mechanical Engineers (ASME) codes, which govern NDT in the oil and gas industry, require that technicians be proficient at estimating flaw heights and characterizing flaws. As a result, it was easier to recruit NDT technicians with the skills needed to participate in this study from the oil and gas industry than from the structural industry. While the NDT technicians had these requisite skills, we found that most technicians in the oil and gas industry were unable to support TFM/FMC and TOFD. As a result, the research team was unable to evaluate those techniques during the round-robin because the technicians were either unfamiliar with the techniques or did not have the equipment needed to implement those techniques. However, it is noted that estimates of flaw length and height obtained using TFM/FMC during the Level III validation of procedures were less uncertain than those obtained using PAUT and TOFD. As the TFM/FMC technique becomes more widely available, it should be further investigated as an NDT technique with potential application to HSS.

6.3 Technician Training and Prequalification

The results of technician prequalification demonstrated that information about an NDT technician's certifications and experience cannot be used to predict that technician's performance with respect to detecting and sizing flaws. There are several reasons for this. As noted, training and testing of NDT technicians varies across states and industries. Another reason is that, in the United States, NDT technicians are tested and certified by their employers. While the standards for certification may be written in the codes governing NDT testing in the industries served by that employer, the rigor with which those standards are applied may vary from one employer to another. There is also a potential conflict of interest. Employers who have invested in the training of an NDT technician have an incentive to put that technician to work and generate profits.

An NDT technician's ability to detect and size flaws must be established through performance qualification, which involves a series of practical and written tests to verify that the NDT technician has the requisite skills to conduct UT on HSS. A minimal amount of performance qualification was employed in this study to verify the skills of NDT technicians participating in round-robin experiments. This was sufficient to eliminate unqualified technicians, including those who were not able to calibrate their instruments or otherwise demonstrate success applying flaw detection and sizing techniques. However, it was not sufficient to distinguish those technicians who were marginal or merely adequate from those who were good. These results demonstrate that a much more extensive performance qualification process will be needed to identify NDT technicians who have the skills required to reliably detect and size flaws in HSS. The capabilities of NDT technicians on the practical portion of the performance qualification test should be evaluated in terms of three metrics: TPR, the ability to size flaws within tolerances (i.e., FST), and PPV. Our results showed that these metrics were not highly correlated with one another, meaning that technicians may rate highly in terms of one metric but not another. Therefore, each metric represents one aspect of overall skill, and all three aspects of skill should be evaluated in performance qualification.

Test specimens used in performance qualification should reflect the joint geometries and weld configurations encountered in HSS. These geometries and weld configurations are often unique and more complex than those found in other structures. For this study, the test specimens used in performance qualification were part of an off-the-shelf kit designed for NDT testing and training in the bridge industry (i.e., the AWS specimens). These specimens have simpler geometries than those encountered in HSS, the welds are ground flush, and the number and distribution of flaws in each specimen is predictable. Our results showed that, during round-robin testing, the TPR in AWS specimens were not correlated with the TPR in ERDC specimens. This suggests that the ability to detect flaws in specimens with simpler geometry is not predictive of the ability to detect flaws in more complex geometries. Round-robin test results also showed that the FST and the PPV in AWS specimens were weakly correlated with those in the ERDC specimens. This suggests that the ability to size flaws in AWS specimens and avoid making false calls in AWS geometries may be somewhat predictive of those abilities in more complex geometries. While this report recommends conducting performance qualification using specimens specifically designed to represent HSS geometries, the greater

priority is to conduct performance qualification; off-the-shelf specimens should be used if those are the only ones available. It is also noted that Annex O of AWS D1.1 (AWS 2020c) requires that a mockup of the specific joint in question be used to qualify the technique and the equipment as well as the technician.

In general, NDT technicians from the structural industry will require additional training before they can participate in performance qualification or be employed to detect and size flaws for FFS analysis of HSS. The agency should make the training available to NDT technicians who are interested in working on HSS. The program should include training on the three recommended UT techniques and on flaw characterization and height sizing methods. It should also identify and remedy any other individual weaknesses. The technician training program and performance qualification should be coordinated. Training should be done by an independent organization that is specifically authorized to do so by the agency. Performance qualification tests should be administered by the agency so that requisite standards are maintained and should be conducted at a central location that is specifically equipped for that purpose and where the test specimens can be stored. Technicians should travel to the site for performance qualification and should bring their own equipment. Instruments should not be shared by two or more technicians because calibration settings and scan results can be stored by one technician and later retrieved by another.

Industry should bear the costs of training and performance qualification up front and recover their costs by providing NDT services to the HSS market. This will help ensure that industry has no incentive to send unqualified technicians to performance qualification events. The need for this arrangement was clearly demonstrated when technicians were paid to participate in performance qualification for round-robin testing. There were numerous cases in which technicians arrived unprepared or unable to participate. In some cases, employers had not informed their technicians of the purpose and objectives of performance qualification. In other cases, technicians did not bring their own equipment despite being told to do so. In another case, the technician did not know how to calibrate their instrument and was unable to begin the practical exam. This was despite explicit instructions stating that each technician would be required to calibrate their instrument. Technicians were also asked to provide information documenting their certifications and experience prior to performance qualification. This information was difficult to compile before

performance qualification. This information was obtained through interviews at the performance qualification site and the documents compiled before round-robin testing.

6.4 POD

Logistic regression models were fit to round-robin test results. These results demonstrated that flaw dimensions are an important factor in determining POD, but they also demonstrated that technician skill and a variety of flaw and specimen characteristics other than flaw dimensions can also influence POD. In particular, a strong positive relationship was established with technician TPR. This demonstrates the importance of technician skill as a factor in determining whether or not a flaw is detected or missed. Several observations demonstrated the validity of the parameter estimates and other conclusions from these logistic regressions. The parameter estimates were stable and remained significant when models were fit to subsets of the data and when independent variables were added or removed from the regression. With the exception of the parameter estimate for length in the volumetric flaw dataset, the sign of significant parameter estimates mirrored the expectations of an ASNT Level III NDT technician. The concordance and the area under the ROC curves demonstrated that the models accurately predicted which flaws were more likely to be detected and which flaws were more likely to be missed in a large fraction of cases, and these statistics remained consistent when models were fit to subsets of the data. While these logistic regression results appear to describe the data to which they are fit, one should be careful about generalizing beyond those data.

Logistic regression results can be extrapolated beyond the data to which they were fit when the external validity of those models has been established. This is demonstrated, for example, by showing that the results of the logistic regression can be replicated on independent data sets. The dataset developed for this study was based on an experimental design that was sparsely populated—meaning there were a limited number of specimens and flaws available for testing and a large number of flaw and specimen characteristics were varied—and there were a limited number of opportunities to scan each flaw. There were also differences in the logistic regression results fit to the planar and volumetric datasets that are difficult to explain, such as the negative parameter estimate on length for volumetric flaws. A more robust experimental design and a larger number of

observations would be needed to support generalizations about the application of PAUT to HSS. However, the results of this study demonstrate that technician skill, specimen geometry, and flaw characteristics other than flaw dimensions can have a significant influence on POD.

While a more robust experimental design and a larger number of observations may be needed to fully estimate what effect these variables have on POD, these results clearly demonstrated that variables other than flaw length and height are also important factors to consider when evaluating the results of NDT on HSS. HSS managers should know the technician's track record of performance and should apply a higher level of scrutiny to PAUT test results in T-joints, corner joints, skewed joints, and thicker plates. While this study focused on the influence of technician skill, specimen geometry, and flaw characteristics, there are a variety of other factors that can also influence POD. These include such things as environmental conditions, time constraints imposed on inspections, technician personality, and fatigue. These factors were not considered in this study.

6.5 Reliability of Indications

True indications are those that can be associated with an actual flaw. False indications are those that cannot. False indications arise routinely in UT results. They are caused by some condition other than a flaw, such as interference from the transducer, the surface of the specimen, mode conversion of the sound beam, or geometry. Knowing the conditions under which false indications are more likely to occur is useful because it enables HSS managers to identify NDT test results that may require additional validation. Logistic regressions are used to estimate the probability that an indication is false and to determine what factors influence that probability. The factors potentially influencing the reliability of an indication are the observable specimen characteristics, the unobservable flaw characteristics, and the skill with which a technician can distinguish between true and false indications.

The reliability of indications was modeled by fitting a logistic regression equation to data on all indications reported during the round-robin experiment. FP indications were distinguished from TP indications using the acceptance algorithm described in Chapter 3. The dependent variable was assigned a value of 1 if it was a TP indication and a value of 0 (zero) otherwise. When technicians report a flaw, they describe its location and

dimensions and characterize its type or category. Unless known otherwise, these factors are unobservable. The observable characteristics that may influence the reliability of an indication include specimen geometry, joint skew, and plate thickness. Knowing the specimen characteristics and reported flaw characteristics that are associated with false indications may help HSS managers identify indications that require additional or independent review.

Approximately 90% of reported indications (i.e., calls) were associated with actual flaws (i.e., they were positive predictions). The greater the flaw length reported by the technician, the more likely it was that the indication was associated with an actual flaw. The greater the aspect of the flaw reported by the technician, the less likely it was that the indication was associated with an actual flaw. Most false calls can be attributed to the influence of joint geometry. However, results showed that (1) indications reported as LOF were less likely to be actual flaws than those reported as belonging to other flaw categories, and (2) indications reported as LOP were more likely to be actual flaws than those reported as belonging to other flaw categories. False calls were 67.5% more likely in ERDC specimens than in AWS specimens. This may be attributed to the more complex geometries represented by ERDC specimens.

6.6 Uncertainty in Flaw Size Estimation

The term flaw size refers to the collective dimensions of a flaw. All methods of estimating flaw dimensions are inherently uncertain. The amount of error depends on many factors, including technician skill, flaw characteristics, specimen geometry, and others that are much harder to replicate and quantify, such as fatigue or inspection environment. This study quantified uncertainty in the ability of NDT technicians to estimate flaw length and height. During round-robin testing, technicians determined the terminal locations of each flaw in each dimension using the decibel drop method or diffraction techniques and used that information to estimate the length and height of each flaw. The research team calculated a proportional error term as the ratio of estimated to actual length and height. Lognormal probability distributions were fit to the proportional error term, and partial safety factors were derived from the lower 95% confidence bound. The partial safety factors can be used in FFS analysis to estimate an upper bound on flaw length and height. The upper bound is calculated by multiplying an estimate of flaw length or height by the partial safety factor to

estimate the length or height below which 95% of estimates will fall. This provides a conservative approach to using estimates of flaw length or height from PAUT in FSS evaluations.

Based on the results of this study, partial safety factors of 2.0 for length and 3.0 for height are recommended when applying length and height estimates from PAUT to FFS analysis. These partial safety factors account for the variability over all technicians, joint geometries, and flaw categories. Describing potential error as proportional to flaw length or height reflects the observation that larger errors are associated with larger flaws. For comparison, Table T.1 of BS7910:2013+A1:2015 (BSI 2015) reported that focused phased array can be used to estimate flaw length with an accuracy of ±7 mm and flaw height with an accuracy of ±1.5 mm. These absolute errors would be consistent with the recommended partial safety factors for a flaw length of 0.55 in. (14 mm) and a flaw height of 0.17 in. (4.5 mm). However, the absolute errors reported in BS7910:2013+A1:2015 (BSI 2015) would tend to understate the amount of uncertainty in larger flaw size estimates.

When our data were separated by flaw and specimen characteristics, partial safety factors were largely consistent with those recommended in the preceding paragraph. For example, when distributions were fit to the ratio of estimated to actual height for different joint geometries, most estimates of partial safety factors were in the range of 2.7–3.03, which is very similar to the recommendation of 3.0. However, the partial safety factor for height in skewed T-joints was much higher: 7.69. Although estimating flaw height in skewed T-joints is generally expected to be more difficult than in some other types of joint geometries, this value seems particularly high. Because there were a relatively small number of observations available to fit this distribution, more study is recommended to validate this result.

There was a large amount of variability in partial safety factors that were calculated for individual NDT technicians. Estimates varied from 1.37 to 2.93 for length and from 2.14 to 4.64 for height. Ten of the 18 technicians had partial safety factors greater than the recommended values for length, height, or both. Interestingly, there was no correlation in the partial safety factors for length and height, and there was only one technician with partial safety factors that exceeded the recommended levels for both length and height. This suggests that an NDT technician's ability to estimate flaw

length is not indicative of that technician's ability to estimate flaw height, and vice versa.

While the proposed approach to using PAUT estimates of length and height is conservative, it should also be recognized that these partial safety factors were derived from estimates made under ideal conditions. Technicians were operating without time constraints and in ideal testing conditions (i.e., indoors, in a climate-controlled environment with good lighting) and may have been more careful because they knew that their work was being evaluated. The stresses and constraints of working in the field, including poor weather, time constraints, and limited access to welds, may justify additional safety factors. In addition, some technicians exhibit more skill than others when sizing flaws, and it may be advisable to take this information into account. For example, if it is believed that a technician's skill level may be at the lower end of the spectrum, additional safety factors may be warranted.

6.7 Flaw Characterization

Flaw characterization is not presently required within the structural NDT industry, and technicians from the structural industry do not generally practice flaw characterization. However, information about the character of detected flaws is an important input to FFS analysis of HSS. The technicians who participated in round-robin testing for this study came from the oil and gas industry. NDT testing in the oil and gas industry is governed by ASME codes that require flaw characterization. For FFS evaluation, sharp, planar discontinuities, such as cracks, are the most critical because they act to intensify stress and are more likely to propagate in the member. Therefore, all flaws must be conservatively considered to be planar unless the flaw type can be reliably characterized otherwise.

This study found that NDT technicians from the oil and gas industry were able to accurately characterize flaws as planar, volumetric, or laminar approximately 80% of the time. Planar flaws were accurately characterized 88% of the time. Volumetric flaws were accurately characterized 55.3% of the time. Laminar flaws were accurately characterized 80% of the time. The low level of accuracy for volumetric flaws reflects the tendency to classify flaws as planar unless they can be reliably classified as volumetric. While it is most important that technicians be able to distinguish between planar, volumetric, and laminar flaws, this study also assessed flaw

characterization accuracy using classification systems with 6 and 10 flaw subcategories. Flaw characterization accuracy decreased to 59.2% for the classification system with 6 subcategories and to 50.7% for the classification system with 10 levels. These results demonstrate that classification accuracy decreases as the number of flaw categories increases.

6.8 Verification Sampling

The results of this study demonstrate that NDT technicians can frequently miss flaws, false indications are common, there is a large amount of uncertainty in flaw size estimates, and the accuracy with which technicians can characterize flaws is limited. These conclusions are based on data collected during round-robin experiments carried out in an ideal testing environment. There are additional challenges associated with field testing that suggest that it may be difficult for NDT technicians to maintain the levels of performance documented in this study during field testing. Therefore, it is recommended that a sample of each NDT technician's work be verified by an independent third party who has previously demonstrated a high level of performance as an NDT technician. The sample should include both positive and negative test results. Because the results of this study indicate that FN results are more likely than FP results, it may be that negative test results should be verified more frequently than positive results. The proportion of NDT test results subject to verification sampling should also reflect the presence or absence of factors that are known to have a negative influence on POD, the reliability of indications, and flaw size estimates.

6.9 Summary of Recommendations

The following list of recommendations is provided to summarize the conclusions of this report:

1. Prior to engagement for work on HSS, all NDT technicians should be required to pass an independent, authorized examination to certify their ability to apply the requisite UT techniques to existing HSS. The practical portion of this examination should be conducted using weld specimens that represent the joint geometries encountered in existing HSS. The requirements for certification should include minimum standards of performance with respect to detection, sizing, and flaw characterization.
2. A training program that is independent of and approved by USACE should be available to NDT technicians to develop the requisite skills and

experience to work on HSS. Training should include advanced UT techniques (i.e., PAUT, TFM/FMC, and TOFD), height sizing, flaw characterization, and coordinate systems for HSS.

3. NDT procedures for HSS should incorporate the following elements: (1) manual rastering, (2) line scans from multiple index offsets, (3) scanning from all faces, and (4) smaller probes.
4. An NDT technician with known capabilities with respect to detecting, sizing, and characterizing flaws should verify a sample of positive and negative NDT test results in the field (i.e., verification sampling). This applies to both in-house labor and contractors.
5. The proportion of NDT test results subject to verification sampling should reflect the presence or absence of factors known to have a negative influence on POD, the reliability of indications, and flaw size estimates.
6. Uncertainty in flaw size estimates can be accounted for in FFS using a partial safety factor of approximately 2 for flaw length and approximately 3 for flaw height. Depending upon the situation, additional margins of safety may be warranted. These factors of safety were not calibrated to provide a target level of reliability. To achieve this, the safety factors for load and resistance effects must be calibrated simultaneously for HSS. When the FFS evaluation indicates marginal acceptance or rejection or there is uncertainty in the input values, a sensitivity analysis should be performed to assess how changes in the uncertain input values would affect the result.
7. Qualified USACE personnel should witness testing in the field to ensure that testing is performed to the code and to ensure that they understand the results.

Each of these recommendations has already been discussed in other parts of this chapter. Readers should consult those sections for more elaborate discussions of these recommendations. These recommendations should be considered for incorporation into future revisions of EM-1110-2-6054 (USACE 2001).

References

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: Wiley.

ASNT (American Society for Nondestructive Testing). 2020a. *ASNT Standard for Qualification and Certification of Nondestructive Testing Personnel*. ANSI/ASNT CP-189. Columbus OH: American Society for Nondestructive Testing.

ASNT (American Society for Nondestructive Testing). 2020b. *Personnel Qualification and Certification in Nondestructive Testing*. ASNT SNT-TC-1A. Columbus, OH: American Society for Nondestructive Testing.

ASTM (ASTM International). 2023. *Standard Terminology For Nondestructive Examinations*. ASTM E1316-23. West Conshohocken, PA: ASTM International.

AWS (American Welding Society). 2015. *Bridge Welding Code*, 6th ed. AASHTO/AWS D1.5M/D1.5:2015. Miami, FL: American Welding Society. https://pubs.aws.org/Download_PDFS/D1.5M-D1.5-2015-PV.pdf.

AWS (American Welding Society). 2016. *Structural Welding Code—Seismic Supplement*. AWS D1.8/D1.8M. Miami, FL: American Welding Society. https://pubs.aws.org/Download_PDFS/D1.8-D1.8M-2016-PV.pdf.

AWS (American Welding Society). 2018. *Bridge Welding Code*, 7th ed. AASHTO/AWS D1.5M/D1.5:2015. Miami, FL: American Welding Society.

AWS (American Welding Society). 2020a. *Bridge Welding Code*, 8th ed. AASHTO/AWS D1.5M/D1.5:2020. Miami, FL: American Welding Society.

AWS (American Welding Society). 2020b. *Standard Welding Terms and Definitions*, 13th ed. AWS A3.0M/A3.0:2020. Miami, FL: American Welding Society.

AWS (American Welding Society). 2020c. *Structural Welding Code—Steel*, 24th ed. AWS D1.1/D1.1M. Miami, FL: American Welding Society. https://pubs.aws.org/Download_PDFS/D1_1_D1_1M_2020_PV.pdf.

Boone, S., J. Cohen, and P. Sauser. 2019. *Testing of In-Service Bridges Using Automated Ultrasonic Testing Methods*. NCHRP IDEA Project 191. Washington, DC: Transportation Research Board. https://onlinepubs.trb.org/onlinepubs/IDEA/FinalReports/HIghway/NCHRP191.pdf.

BSI (British Standards Institute). 2005. *Guide to Methods for Assessing the Acceptability of Flaws in Metallic Structures*. BS 7910:2005. London, UK: British Standards Institution.

BSI (British Standards Institute). 2015. *Guide to Methods for Assessing the Acceptability of Flaws in Metallic Structures*. BS 7910:2013+A1:2015. London, UK: British Standards Institution.

BSI (British Standards Institute). 2019. *Guide to Methods for Assessing the Acceptability of Flaws in Metallic Structures*. BS 7910:2019. London, UK: British Standards Institution.

Burdekin, F. M., and W. Hamour. 2002. *Partial Safety Factors for the SINTAP Procedure*. HSE Offshore Technology Report 2000/020. London, England: Health and Safety Executive. http://www.hse.gov.uk/research/otopdf/2000/oto00020.pdf.

California DOT (California Department of Transportation). 2021. *Notification of California Department of Transportation Qualification Requirement for Ultrasonic Testing Personnel*. Vallejo, CA: CALTRANS. https://dot.ca.gov/-/media/dot-media/programs/engineering/documents/mets/ut-qualification-program-a11y.pdf.

Carboni, M., and S. Cantini. 2012. "A 'Model Assisted Probability of Detection' Approach for Ultrasonic Inspection of Railway Axles." In *Proceedings, 18th World Conference on Nondestructive Testing* 17, 16–20 April, Durban, South Africa. https://www.ndt.net/?id=12744.

Carter, L., and B. McGrath. 2013. "We Know How to Improve Inspection Reliability—Why Don't We Do It?" Paper Presented at the 5th European-American Workshop on Reliability of NDE Proceedings, Berlin, Germany. https://www.ndt.net/article/reliability2013/papers/lecture19.pdf.

Carvalho, A. A., J. M. A. Rebello, R. R. Silva, and L. V. S. Sagrilo. 2006. "Reliability of the Manual and Automatic Ultrasonic Technique in the Detection of Pipe Weld Defects." *Insight* 48 (11): 649–654. https://10.1784/insi.2006.48.11.649.

Cherry, M., and C. E. Knott. 2022. "What is Probability of Detection?" *Materials Evaluation* 80 (12): 24–28. https://doi.org10.32548.2022-me-04324.

Choi, Y. M., D. I. Kang, Y. L. Kim, S. Cho, T. Park, and I. K. Park. 2022. "Reliability Assessment of PAUT Technique in Lieu of RT for Tube Welds in Thermal Power Plant Facilities." *Applied Sciences* 12 (12): 5,867. https://doi.org/10.3390/app12125867.

CNDE (Center for Nondestructive Evaluation). n.d. "About NDT and NDE." *Iowa State University Center for Nondestructive Evaluation*. Accessed July 05, 2023. https://www.nde-ed.org/About/index.xhtml.

Connor, R. J., C. J. Schroeder, B. M. Crowley, G. A. Washer, and P. E. Fish. 2019. *Acceptance Criteria of Complete Joint Penetration Steel Bridge Welds Evaluated Using Enhanced Ultrasonic Methods*. National Cooperative Highway Research Program Research Report No. 908. Washington, DC: The National Academies Press. https://doi.org/10.17226/25494.

Crutzen, S. J. 1985. "PISC Exercises: Looking for Effective and Reliability Inspection Procedures." *Nuclear Engineering and Design* 86 (2): 197–218. https://doi.org/10.1016/0029-5493(85)90223-7.

D'Agostino, A., S. Morrow, C. Franklin, and N. Hughes. 2017. *Review of Human Factors Research in Nondestructive Testing Examination*. Rockville, MD: US Nuclear Regulatory Commission. https://www.nrc.gov/docs/ML1705/ML17059D745.pdf.

Dexter, R. J., H. N. Mahmoud, J. A. Padula, and G. A. Riveros. 2007. *Fitness-for-Purpose Evaluation of Hydraulic Steel Structures*. ERDC TR-07-15. Vicksburg, MS: US Army Engineer Research and Development Center. http://hdl.handle.net/11681/8510.

Ditchburn, R. J., and M. E. Ibrahim. 2009. *Ultrasonic Phased Arrays for the Inspection of Thick-Section Welds*. Defence Science and Technology Office Technical Note DSTO-TN-0911. Canberra, Australia: Australian Government Department of Defence. http://dsto.defence.gov.au/corporate/reports/DSTO-TN-0911.pdf.

DoD. 2009. *Nondestructive Evaluation System Reliability Assessment*. MIL-HDBK-1823-A. Washington DC: United States Department of Defense. https://statistical-engineering.com/wp-content/uploads/2017/10/MIL-HDBK-1823A2009.pdf.

DoD. 2021. *Fabrication of Hydraulic Steel Structures*. Unified Facilities Criteria Guide Specification (UFGS) 05 59 20. Washington, DC: DoD. https://www.wbdg.org/FFC/DOD/UFGS/UFGS%2005%2059%2020.pdf.

Dymkin, G. Y., and V. N. Konshina. 2000. "Systematics of NDT Reliability: A Practical Point of View." In *Proceedings, 15th WCNDT–World Conference on Nondestructive Testing*, 15–21 October, Rome, Italy. https://www.ndt.net/article/wcndt00/papers/idn738/idn738.htm.

ENA (Energy Networks Association). 1983. *Manual Ultrasonic Testing of Welds in Ferritic Steel Sections*. Technical Specification 98-10, Issue 1. London, England: ENA Ltd.

Enkvist, J., A. Edland, and O. Svenson. 2000. *Operator Performance in Non-Destructive Testing: A Study of Operator Performance in a Performance Test*. SKI Report 00:26. Stockholm, Sweden: Swedish Nuclear Power Inspectorate. https://www.stralsakerhetsmyndigheten.se/contentassets/044957fa3e2c4b159eda065e04e09cee/0026-operator-performance-in-non-destructive-testing-a-study-of-operator-performance-in-a-performance-test.

EPRI (Electric Power Research Institute). 2018. *Nondestructive Evaluation: Cast Austenitic Stainless Steel Round-Robin Study Summary of Results, Revision 1*. Report No. 3002010314. Palo Alto, CA: EPRI. https://www.epri.com/research/products/000000003002010314.

Fucsok, F. 1998. "Experiences and Problems of a Manual Ultrasonic Round-Robin Test." In *Proceedings 7th European Conference on Nondestructive Testing*, Copenhagen, Denmark, May 1998. https://www.ndt.net/article/ecndt98/weld/190/190.htm.

Grandt, A. F., Jr. 2011. "Damage Tolerant Design and Nondestructive Inspection—Keys to Aircraft Airworthiness." *Procedia Engineering* 17: 236–246. https://doi.org/10.1016.j.proeng.2011.10.025.

Graybeal, B. A., D. D. Rolander, B. M. Phares, M. E. Moore, and G. A. Washer. 2001. "Reliability and Accuracy of In-Depth Inspection of Highway Bridges." *Transportation Research Record* 1,749 (1): 93–99. https://doi.org/10.3141/1749-14.

Gruber, G. J., and G. M. Light. 2002. "Supplemental Ultrasonic Code Inspection of Structural Weldments." *Journal of Materials in Civil Engineering* 14 (1): 57–61. https://doi.10.1061/(ASCE)0899-1561(2002)14:1(57).

Jacob, R. E., T. L. Moran, A. E. Holmes, A. A. Diaz, and M. S. Prowant. 2018. *Interim Analysis of the EPRI CASS Round Robin Study*. Richland, WA: Pacific Northwest National Laboratory. https://www.nrc.gov/docs/ML1821/ML18219B319.pdf.

Keprate, A., and R. M. C. Ratnayake. 2015. "Probability of Detection as a Metric for Quantifying NDE Capability: The State of the Art." *The Journal of Pipeline Engineering* 14: 199–209.

Knott, C. E., and C. Schubert Kabban. 2022a. "Confidence Interval Comparisons for Probability of Detection on Hit/Miss Data." *Materials Evaluation* 80 (12): 50–65. https://doi.org/10.32548/2022.me-04273.

Knott, C. E., and C. Schubert Kabban. 2022b. "Modern Design and Analysis for Hit/Miss Studies Using Profile Likelihood Ratio Confidence Intervals." *Materials Evaluation* 80 (12): 32–49. https://doi.org/10.32548/2022.me-04272.

Kurz, J. H., A. Jüngert, S. Dugan, G. Dobmann, and C. Boller. 2013. "Reliability Considerations of NDT by Probability of Detection (POD) Determination Using Ultrasound Phased Array." *Engineering Failure Analysis* 35: 609–617. https://doi.org/10.1016/j.engfailanal.2013.06.008.

New York State Department of Labor. 2021. "Certified Ultrasonic Technician." https://statistics.labor.ny.gov/olcny/certified-ultrasonic-technician.shtm.

OECD (Organisation for Economic Co-Operation and Development). 1986. *A Summary of the PISC-II Project*: PISC II Report No. 1. Paris, France: OECD. https://www.oecd-nea.org/upload/docs/application/pdf/2020-01/csni86-117.pdf.

Rummel, W. D. 2004. "Qualification and Validation of the Performance Capability (POD) for Nondestructive Inspection Procedures." In *Proceedings 16th World Conference on Nondestructive Testing*, Montreal, Quebec, Canada, September 2004. https://www.ndt.net/article/wcndt2004/pdf/reliability/619_rummel.pdf.

SAS Institute. n.d. *Logistic Modeling with Categorical Predictors*. Cary, NC: SAS Institute. Accessed July 17, 2023. https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_logistic_examples02.htm.

Schneider, C., and C. R. Bird. 2009. "Reliability of Manually Applied Phased Array Inspection." In *Proceedings, 4th European-American Workshop on Reliability of NDE*, 24–26 June, Berlin, Germany. http://www.ndt.net/index.php?id=8325.

Schultz, M. T., J. H. Milligan, B. E. Skahill, L. E. Campbell, P. W. Sauser, and R. D. Bell. 2023. *Strength and Toughness Inputs to Fitness for Service Analysis of Existing Hydraulic Steel Structures*. ERDC/CHL CHETN-IX-61. Vicksburg, MS: US Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory. http://dx.doi.org/10.21079/11681/46360.

Shaw, R. E., Jr. 2002. "Ultrasonic Testing Procedures, Technician Skills, and Qualifications." *Journal of Materials in Civil Engineering* 14 (1): 62–67. https://doi.org/10.1061/(ASCE)0899-1561(2002)14:1(62).

Singh, R. 2001. "Three Decades of NDT Reliability Assessment and Demonstration." *Materials Evaluation* 59 (7): 856–860.

Spencer, F. W. 1996a. "Visual Inspection Reliability of Transport Aircraft." In *Proceedings of SPIE 2,945, Nondestructive Evaluation of Aging Aircraft, Airports, and Aerospace Hardware*, November 14, 1996, Scottsdale, AZ. http://doi.org/10.1117/12.259089.

Spencer, F. W. 1996b. *Visual Inspection Research Project Report on Benchmark Inspections*. Report No. DOT/FAA/AR-96/65. Washington, DC: Office of Aviation Research. https://apps.dtic.mil/sti/pdfs/ADA321199.pdf.

Spencer, F. W. 2001. "Estimating Probability of Detection Curves from Regression Data." *Materials Evaluation* 59 (7): 866–870.

Stephens H. M. 2000. "NDE Reliability–Human Factors–Basic Considerations." In *Proceedings, 15th World Conference on Nondestructive Testing*, October 2000, Rome, Italy. https://www.ndt.net/article/wcndt00/papers/idn736/idn736.htm.

Swensson, R. G., S. J. Hessel, and P. G. Herman. 1977. "Omissions in Radiology: Faulty Search or Stringent Reporting Criteria?" *Diagnostic Radiology* 123 (3): 563–567. https://doi.org/10.1148/123.3.563.

SwRI and NASA (Southwest Research Institute and NASA Johnson Space Center). 2022a. NASGRO® Fracture Mechanics and Fatigue Crack Growth Analysis Software, v10.0. San Antonio, TX: SwRI.

SwRI and NASA (Southwest Research Institute and NASA Johnson Space Center). 2022b. *NASGRO® Fracture Mechanics and Fatigue Crack Growth Analysis Software Reference Manual, Appendix X–Alternative Failure Criteria*. San Antonio, TX: SwRI.

USACE (US Army Corps of Engineers). 1995. *Periodic Inspection and Continuing Evaluation of Completed Civil Works Structures*. ER 1110-2-100. Washington DC: Department of the Army, US Army Corps of Engineers.

USACE (US Army Corps of Engineers). 2001. *Inspection, Evaluation, and Repair of Hydraulic Steel Structures*. EM-1110-2-6054. Washington DC: Department of the Army, US Army Corps of Engineers. https://www.publications.usace.army.mil/portals/76/publications/engineermanuals/em_1110-2-6054.pdf.

USACE (US Army Corps of Engineers). 2009. *Responsibility for Hydraulic Steel Structures*. ER-1110-2-8157. Washington DC: Department of the Army, US Army Corps of Engineers.

Virkkunen, I., T. Koskinen, and O. Jessen-Juhler. 2021. "Virtual Round Robin—A New Opportunity to Study NDT Reliability." *Nuclear Engineering and Design* 380: 111297. https://doi.org/j.nucengdes.2021.111297.

Appendix A: Examples of Representative Hydraulic Steel Structure (HSS) Joint Geometries

This appendix describes each of the eight representative hydraulic steel structure (HSS) joint geometries that were used as a basis for designing the US Army Engineer Research and Development Center (ERDC) specimens. Each of these joints was labeled (i.e., A through H) and is described in Table A-1. Joints A through E are butt joints, joints F and G are T-joints, and joint H is a corner joint. Figure A-1 through Figure A-9 provide examples of each joint in drawings from representative HSS, which are described in Table A-1. References to the figures and ERDC specimens that represented these joints in the round-robin study are noted in the table.

Table A-1. Descriptions and examples of representative joints used as the basis for US Army Engineer Research and Development Center (ERDC) specimens.

Joint	Description	HSS Example	Figure	Specimens
A	Butt Joint (with or without thickness transition, with web plate attachment)	This joint is similar to the girder flange to diaphragm connection in a miter gate or the flange to gusset plate connection in a sector gate.	A1, A2	ERDC 001
B	Butt Joint (with or without thickness transition, without web plate attachment)	This joint is similar to a splice in a miter gate diagonal or splice in the skin plate of a gate or bulkhead.	A3, A5	ERDC 002, 003
C	Skewed butt joint (adjoining angle between 90 and 180°)	This joint is similar to the flange plate connections in the tapered end section of a miter gate.	A4	ERDC 004
D	Butt joint between plates of different widths (adjoining angle between 0 and 90°)	This joint is similar to the flange to gusset plate connection in a sector gate or the strut bracing connections in a tainter gate.	A5	ERDC 005, 006
E	Butt joint between wide flange beams	This joint is similar to a splice in a wide flange beam in a sector gate or bulkhead.	A7, A8	ERDC 007
F	T-joint (adjoining angle of 90°)	This joint is similar to a lifting lug connection in a gate or bulkhead or the web-to-flange connection in a built-up plate girder.	A1, A6	ERDC 008, 009
G	Skewed T-joint (adjoining angle less than 90)	This joint is similar to the truss connections in a sector gate (flange to flange weld).	A7	ERDC 010
H	Corner joint or skewed corner joint (adjoining angle between 0 and 90°, inclusive)	This joint is similar to the flange plate connections in a sector gate “ear” and end plate connections in a miter gate.	A4, A9	ERDC 011, 012

Figure A-1. Examples of Joint A (US Army Engineer Research and Development Center [ERDC] 001) and Joint F (ERDC 009).

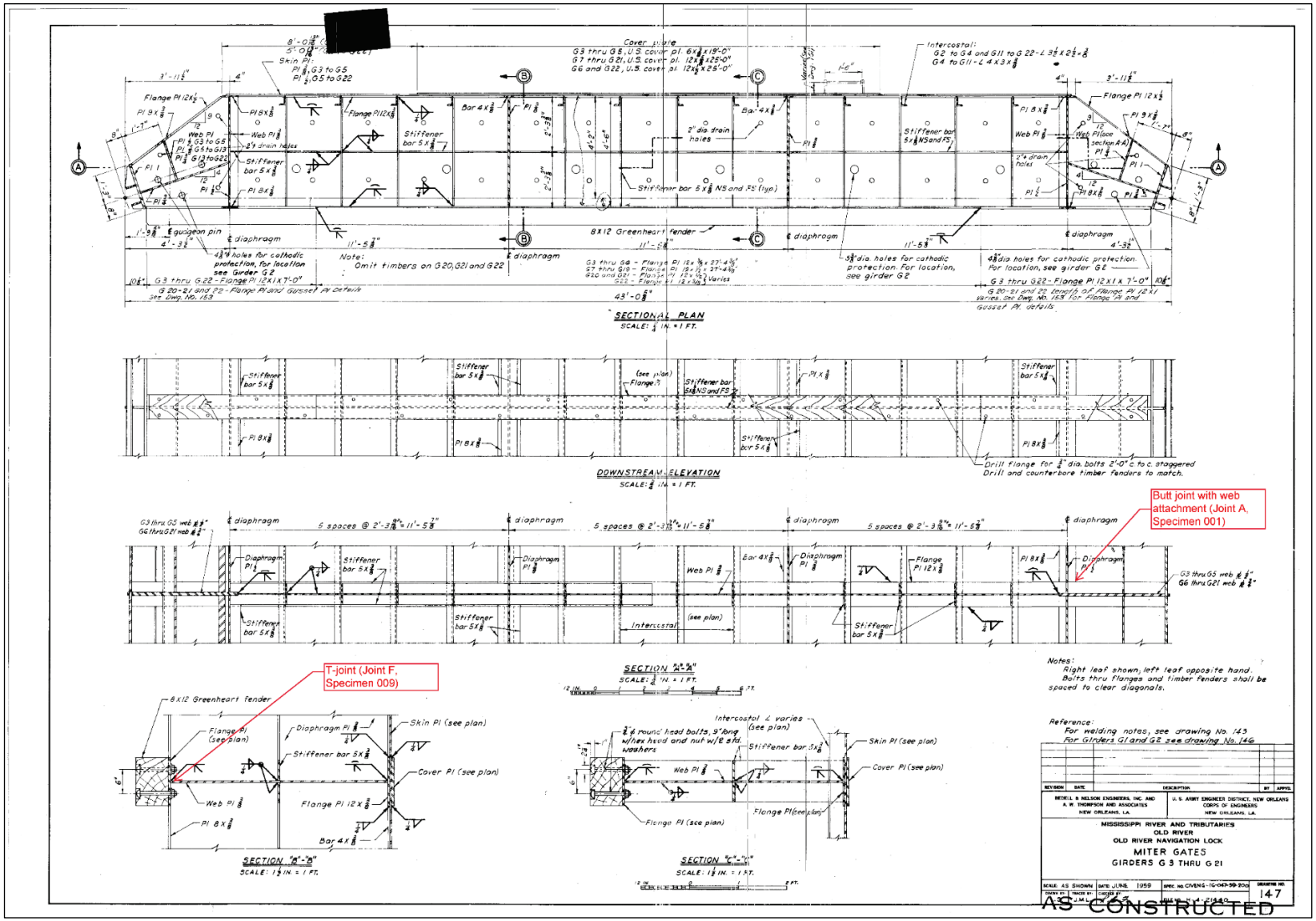


Figure A-2. Example of Joint A (ERDC 001).

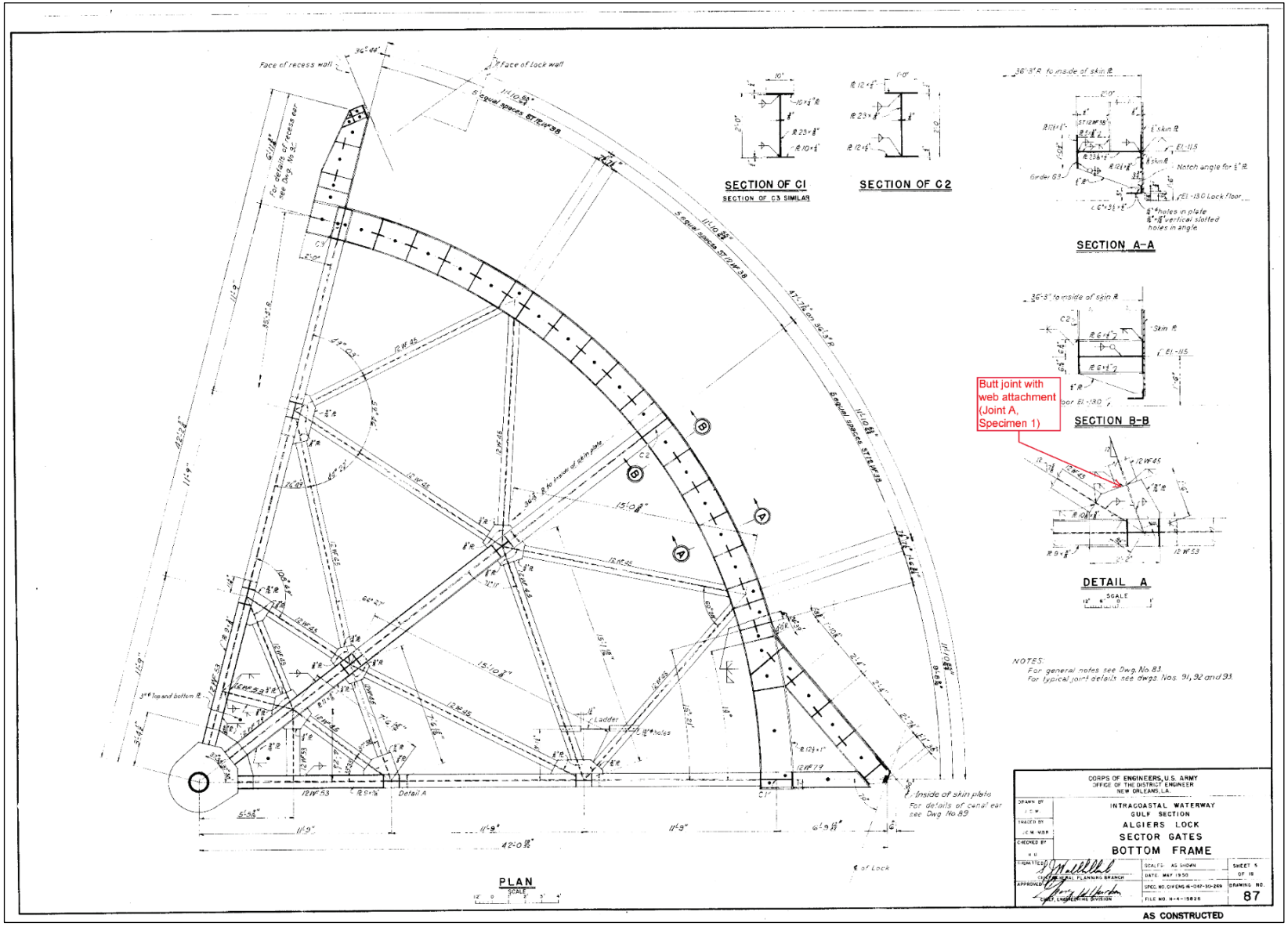


Figure A-3. Example of Joint B (ERDC 002).

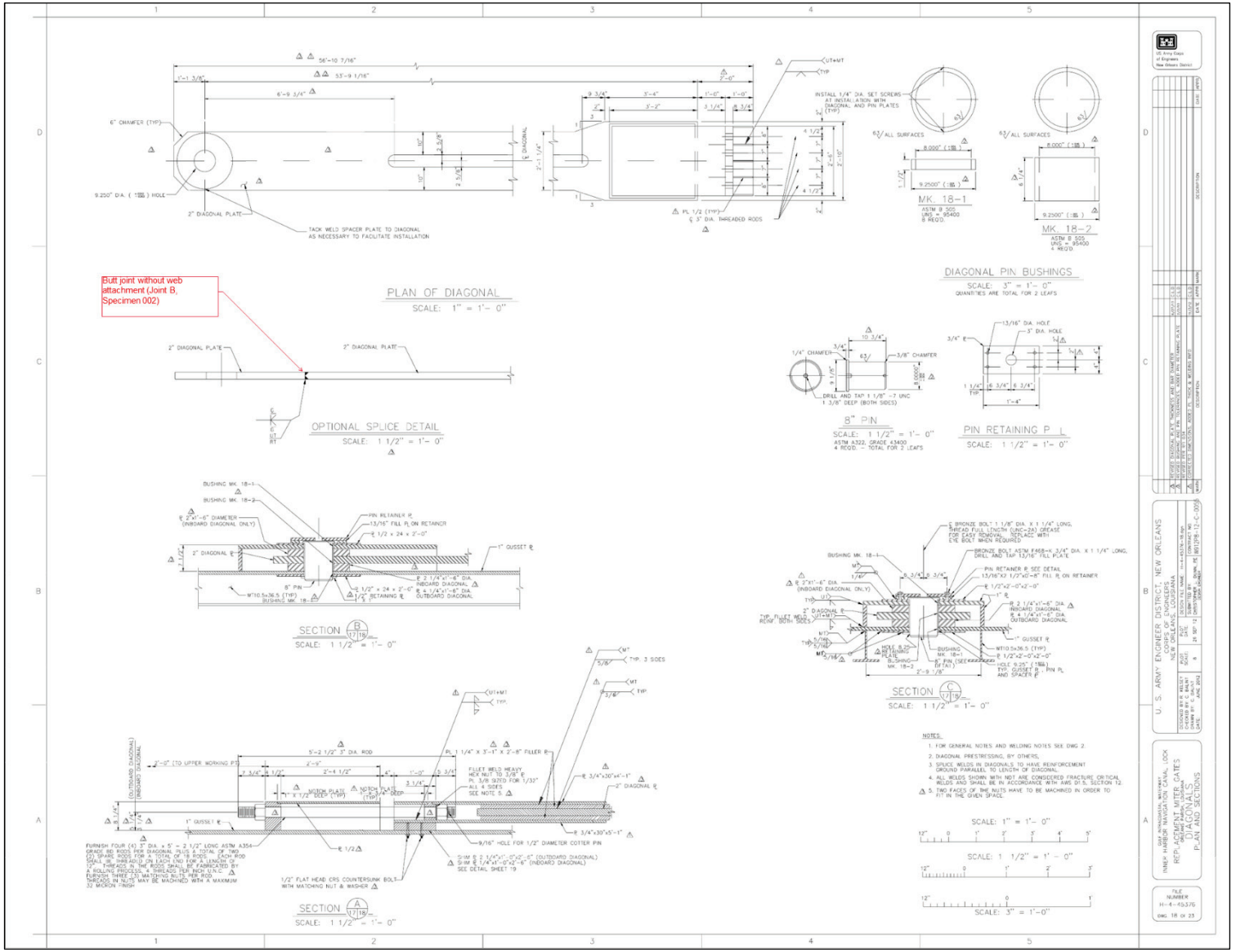


Figure A-4. Examples of Joint C (ERDC 004) and Joint H (ERDC 012).

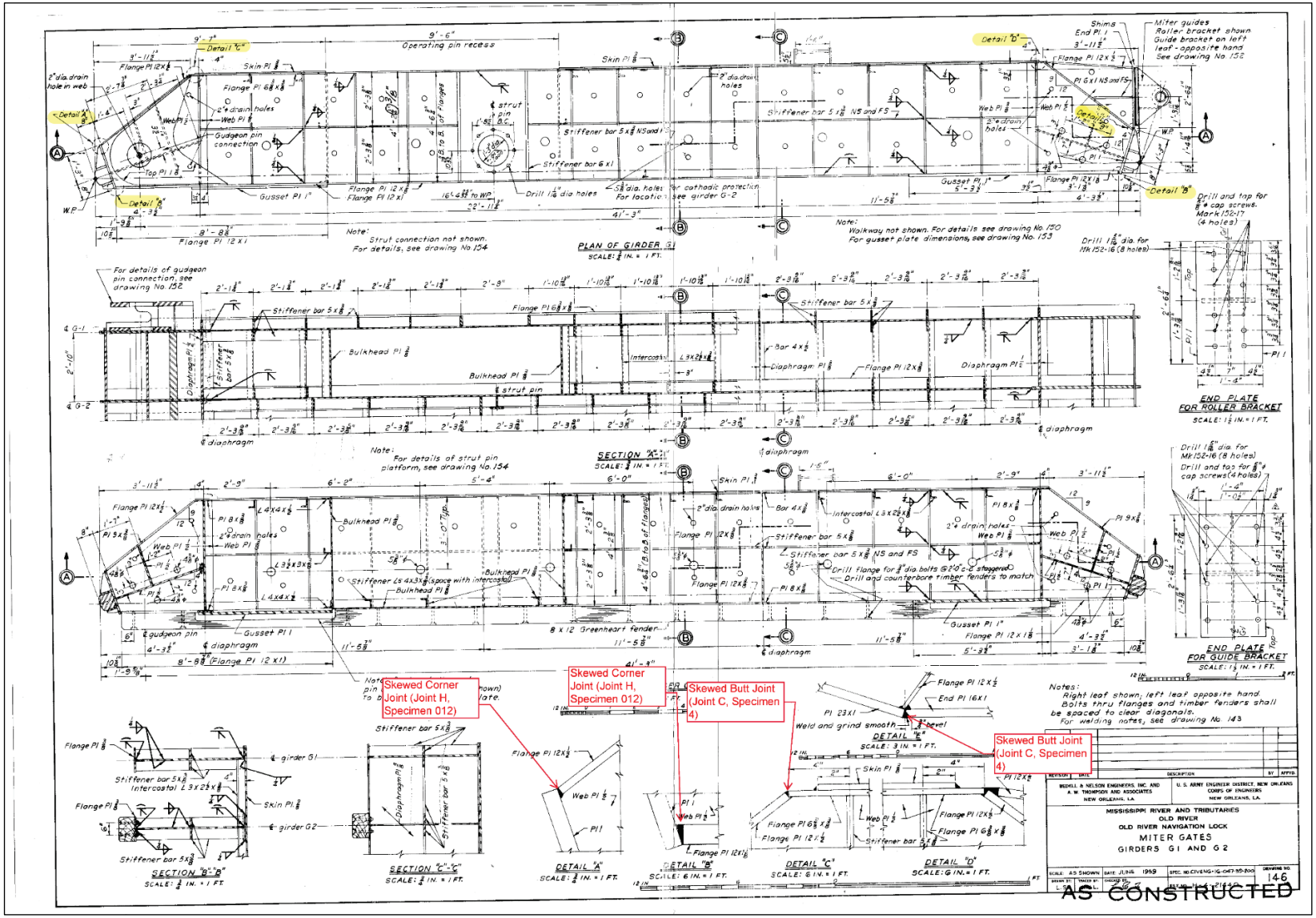


Figure A-5. Examples of Joint D (ERDC 005) and Joint B (ERDC 003).

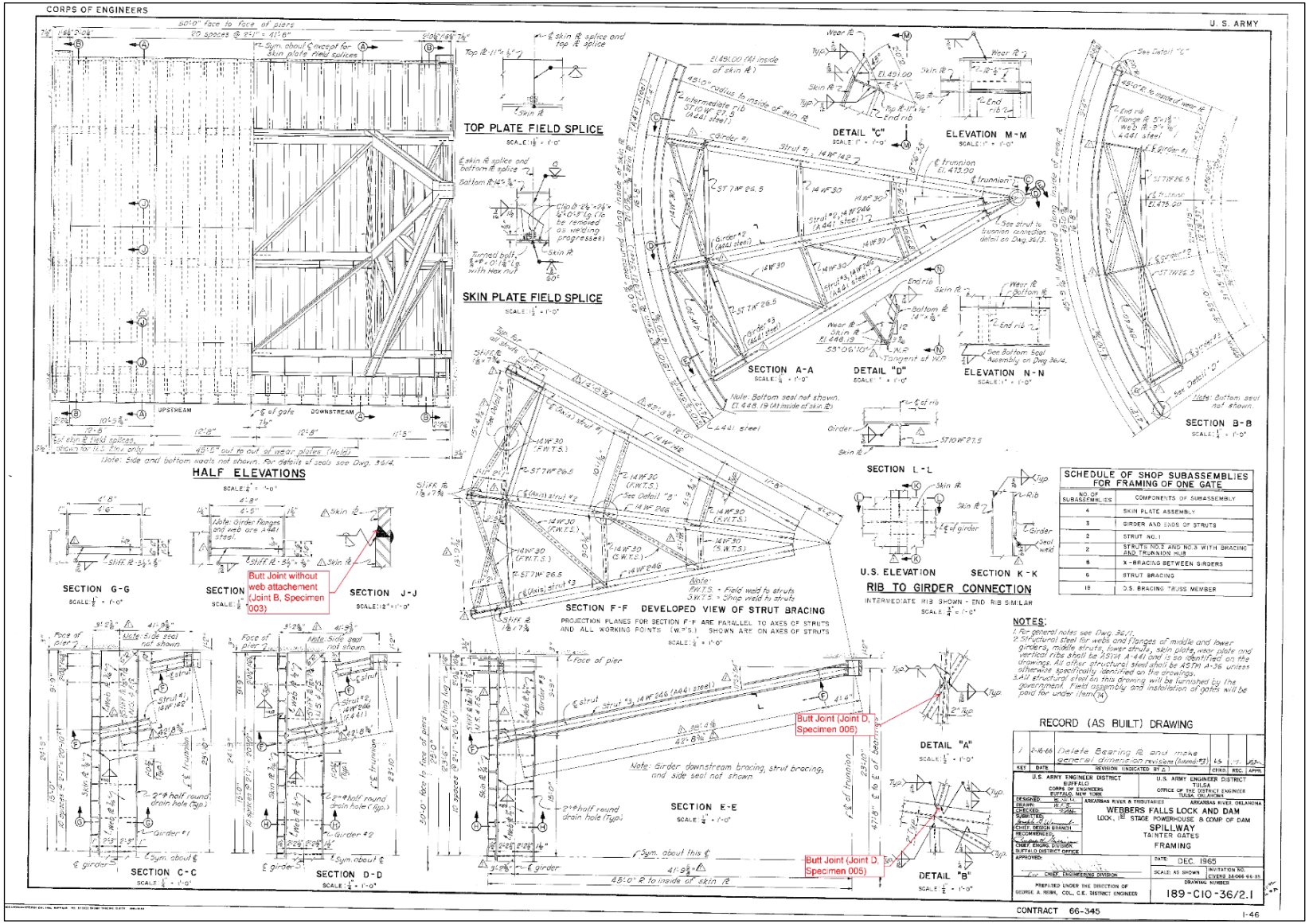


Figure A-6. Example of Joint F (ERDC 008).

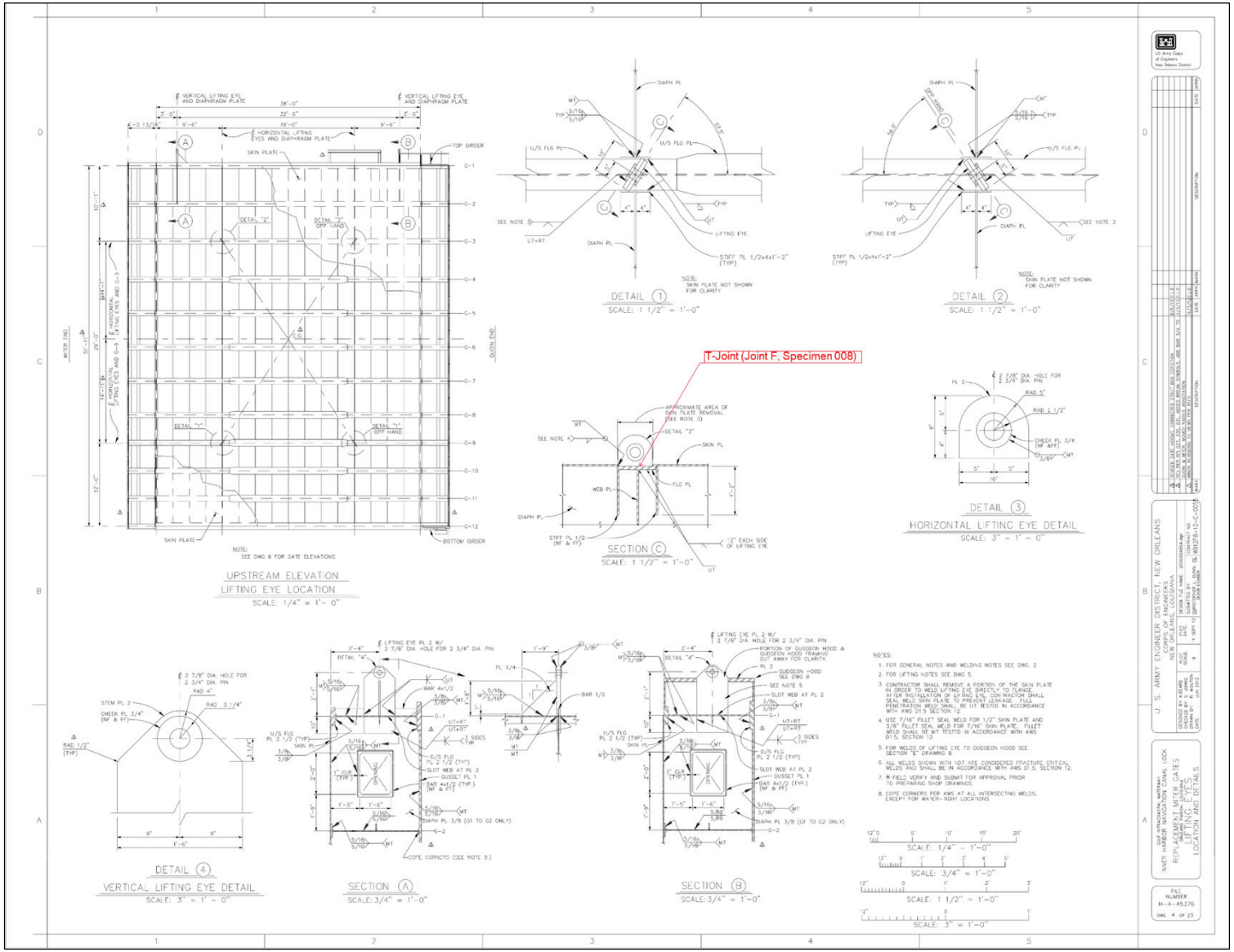


Figure A-7. Examples of Joint E (ERDC 007) and Joint G (ERDC 010).

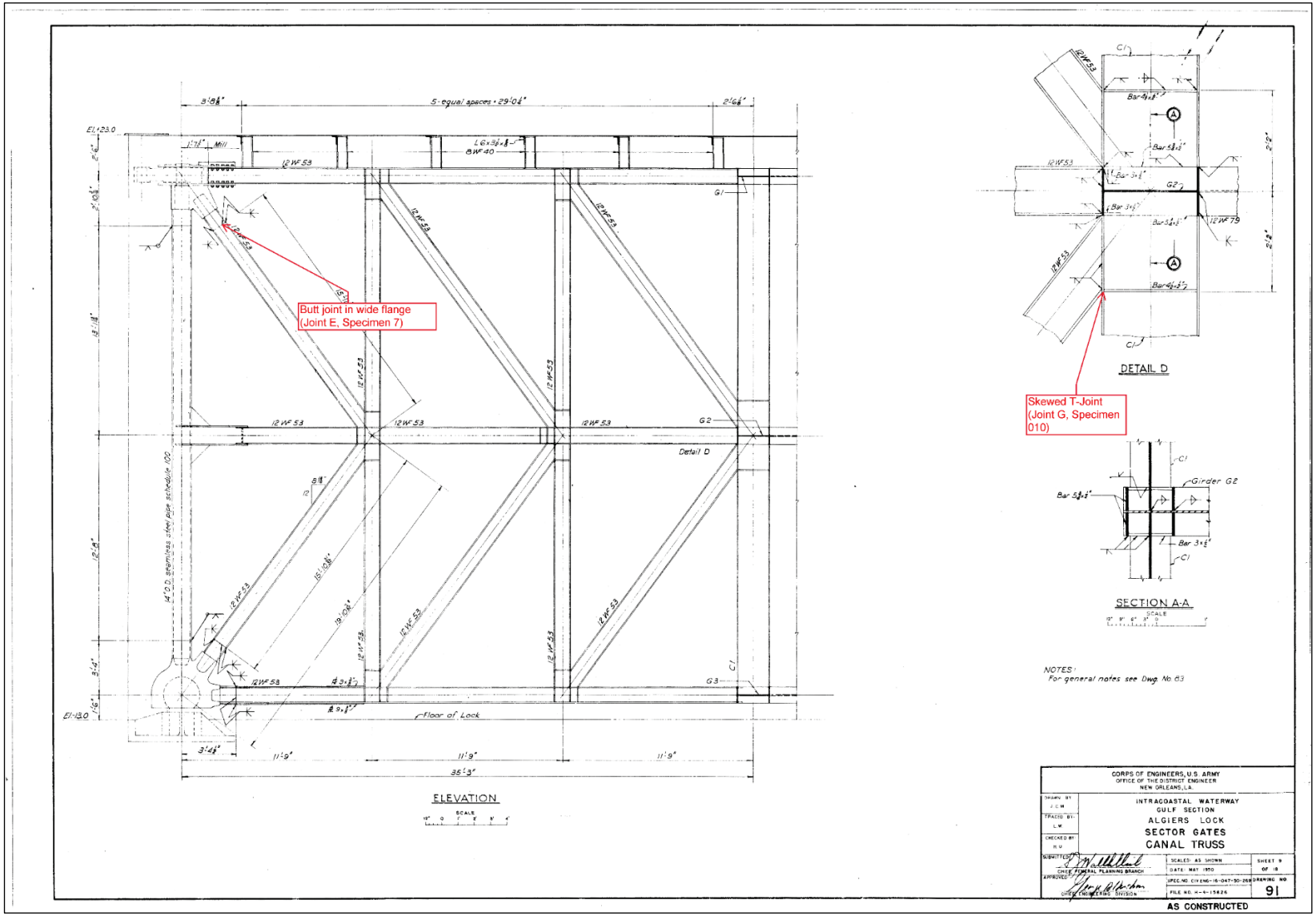


Figure A-8. Example of Joint E (ERDC 007).

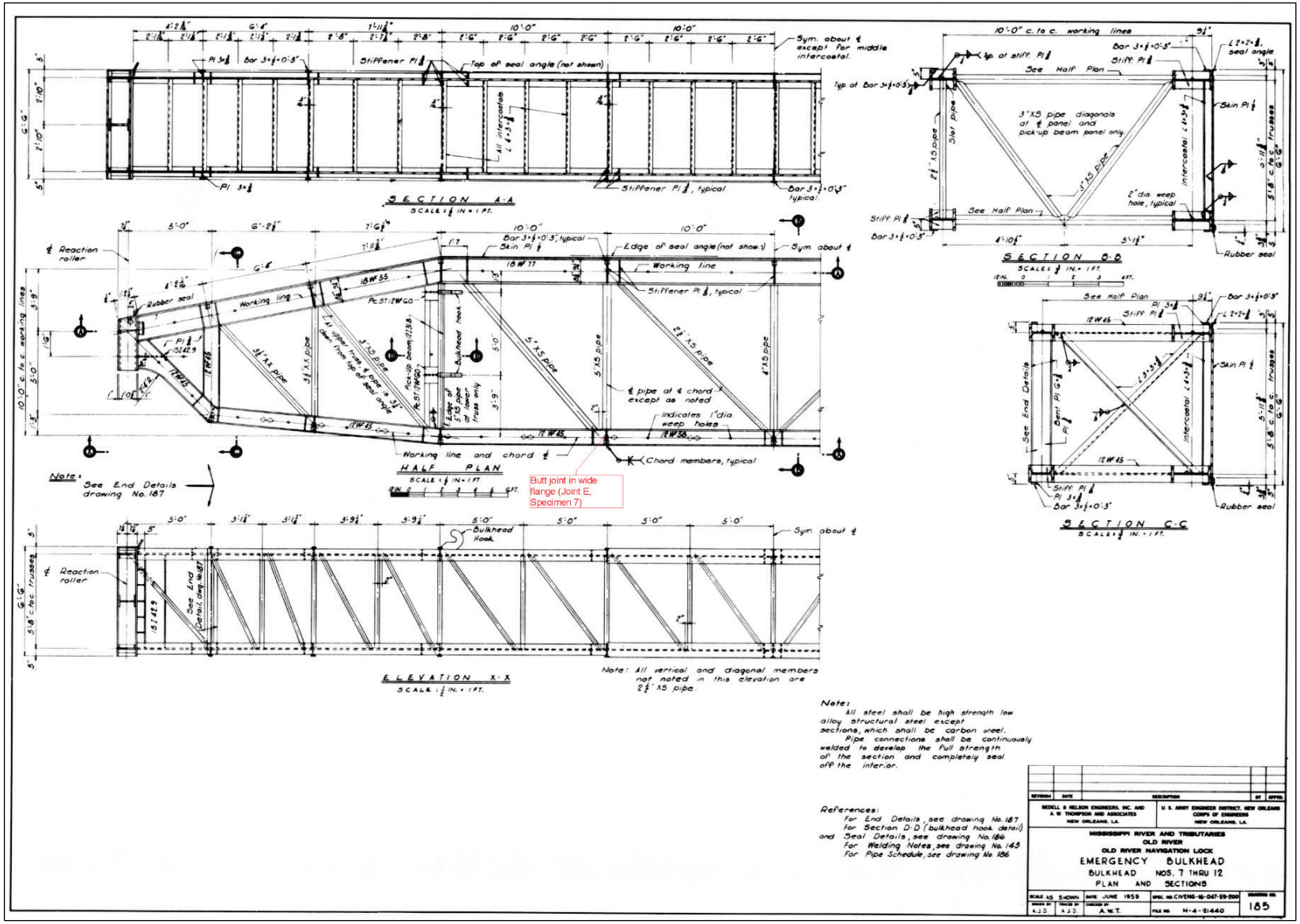
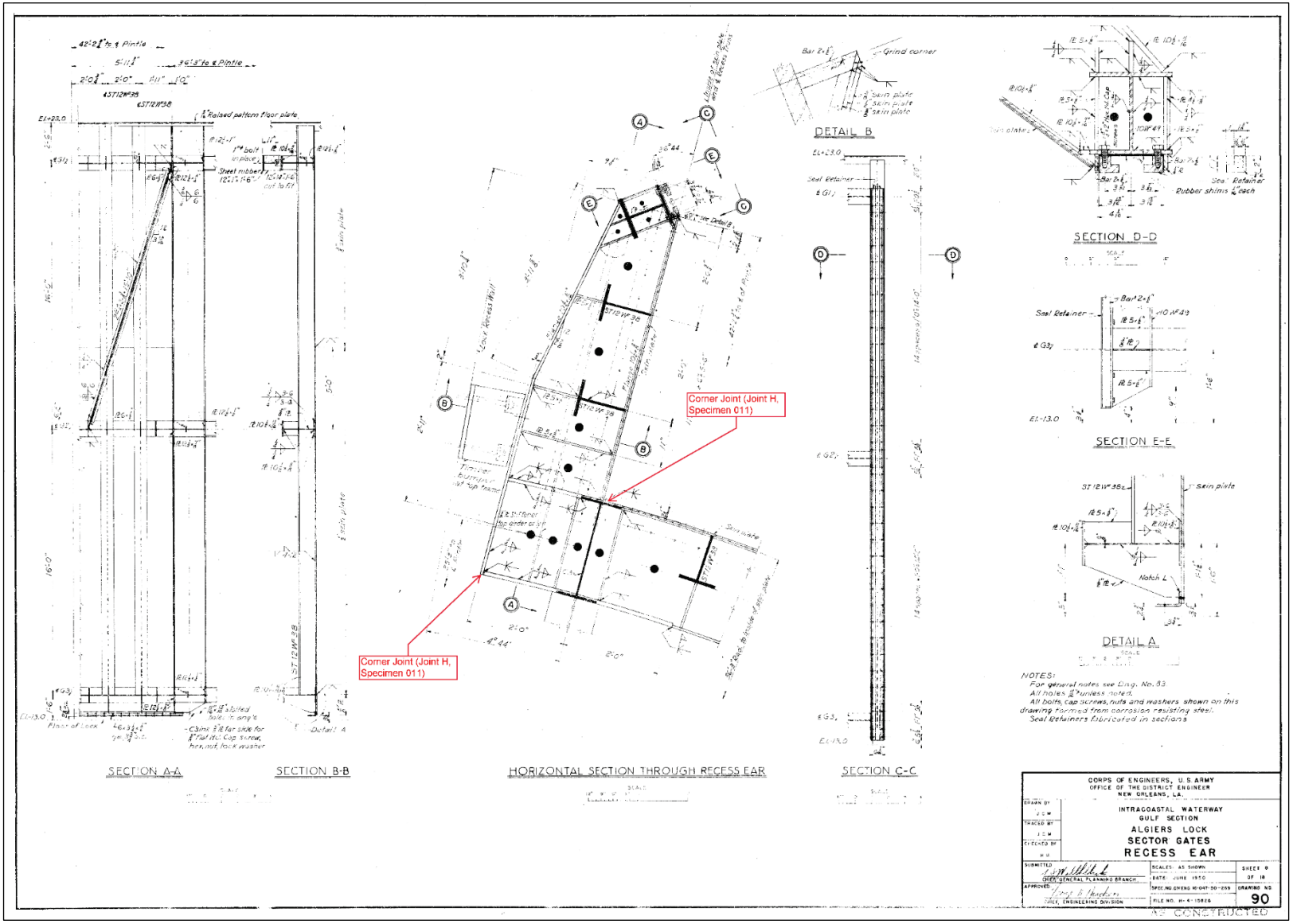


Figure A-9. Examples of Joint H (ERDC 011).



Appendix B: Fabrication Drawings and Photographs of US Army Engineer Research and Development Center (ERDC) Specimens

Figure B-1 through Figure B-24 contain the fabrication drawings and photographs of ERDC specimens used for this research effort.

Figure B-1. Fabrication drawing for ERDC 001.

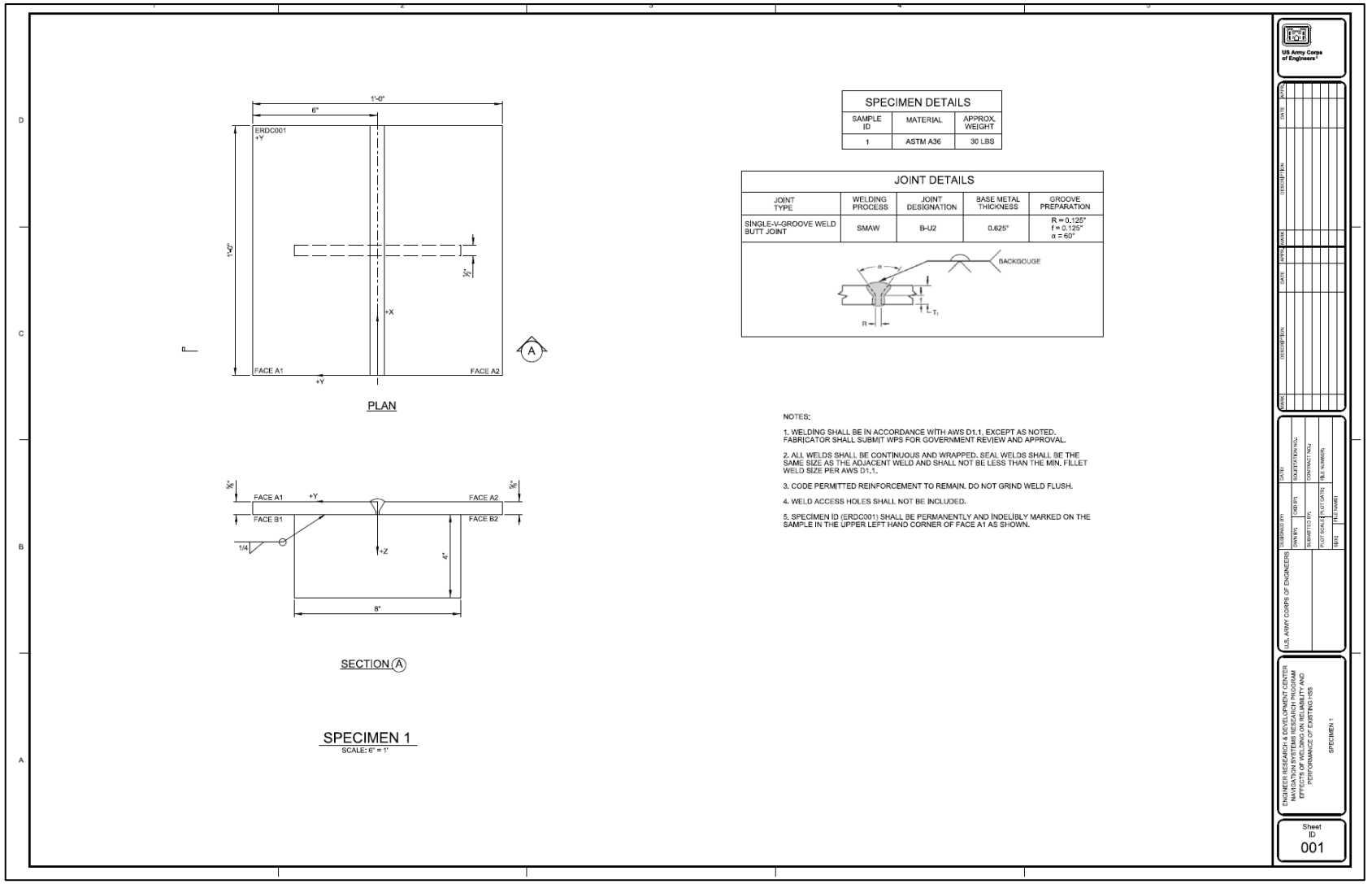


Figure B-2. Photograph of ERDC 001.

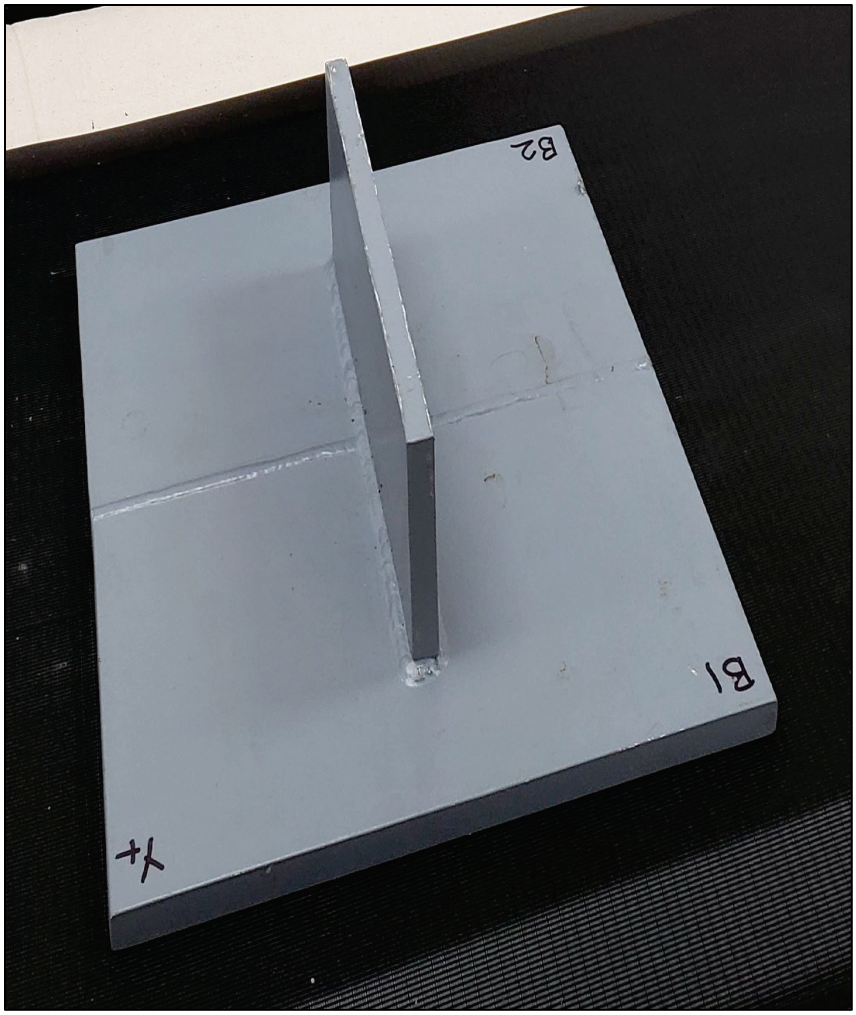


Figure B-3. Fabrication drawing of ERDC 002.

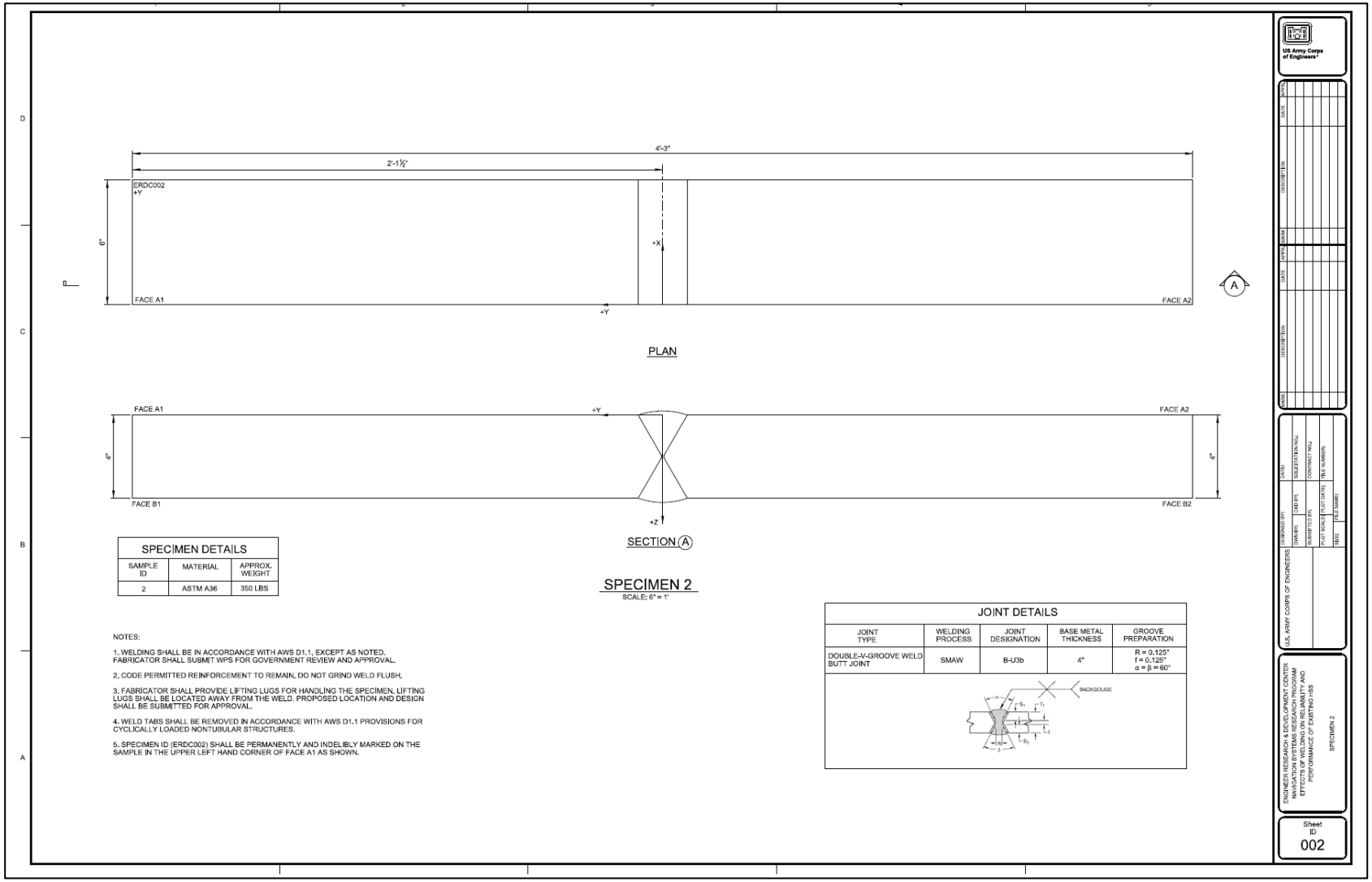


Figure B-4. Photograph of ERDC 002.

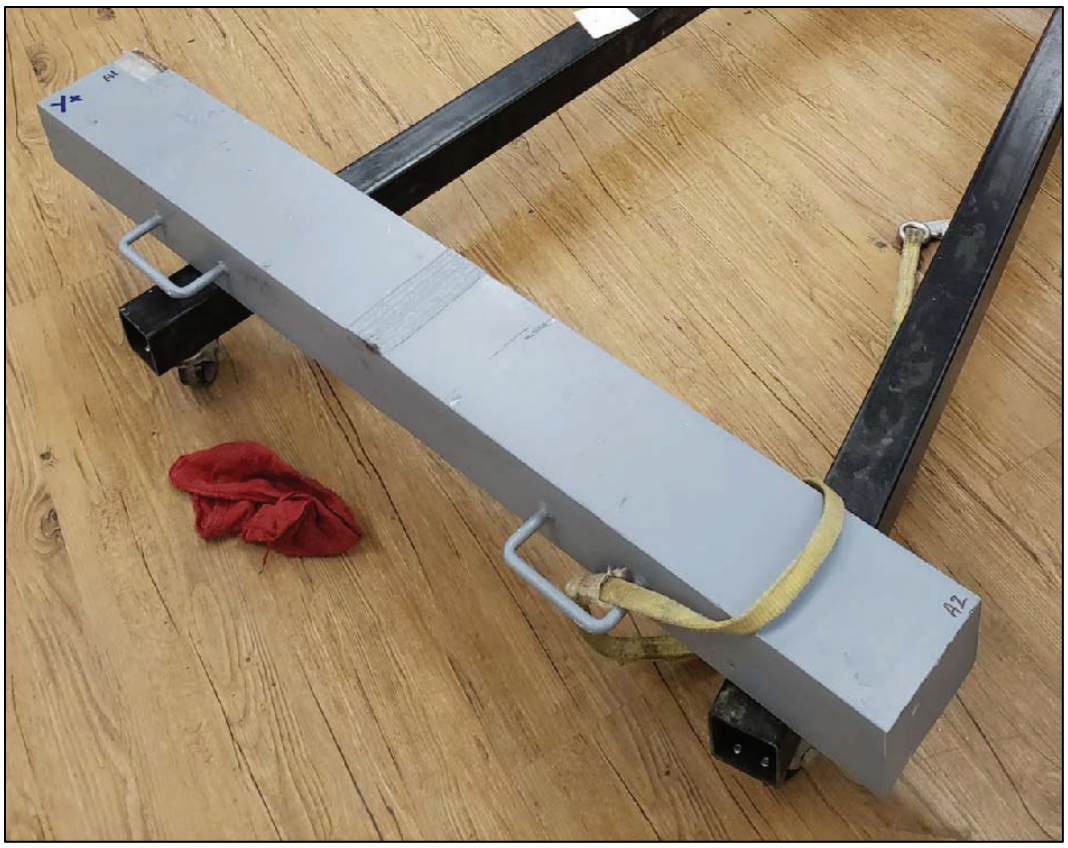


Figure B-5. Fabrication drawing for ERDC 003.

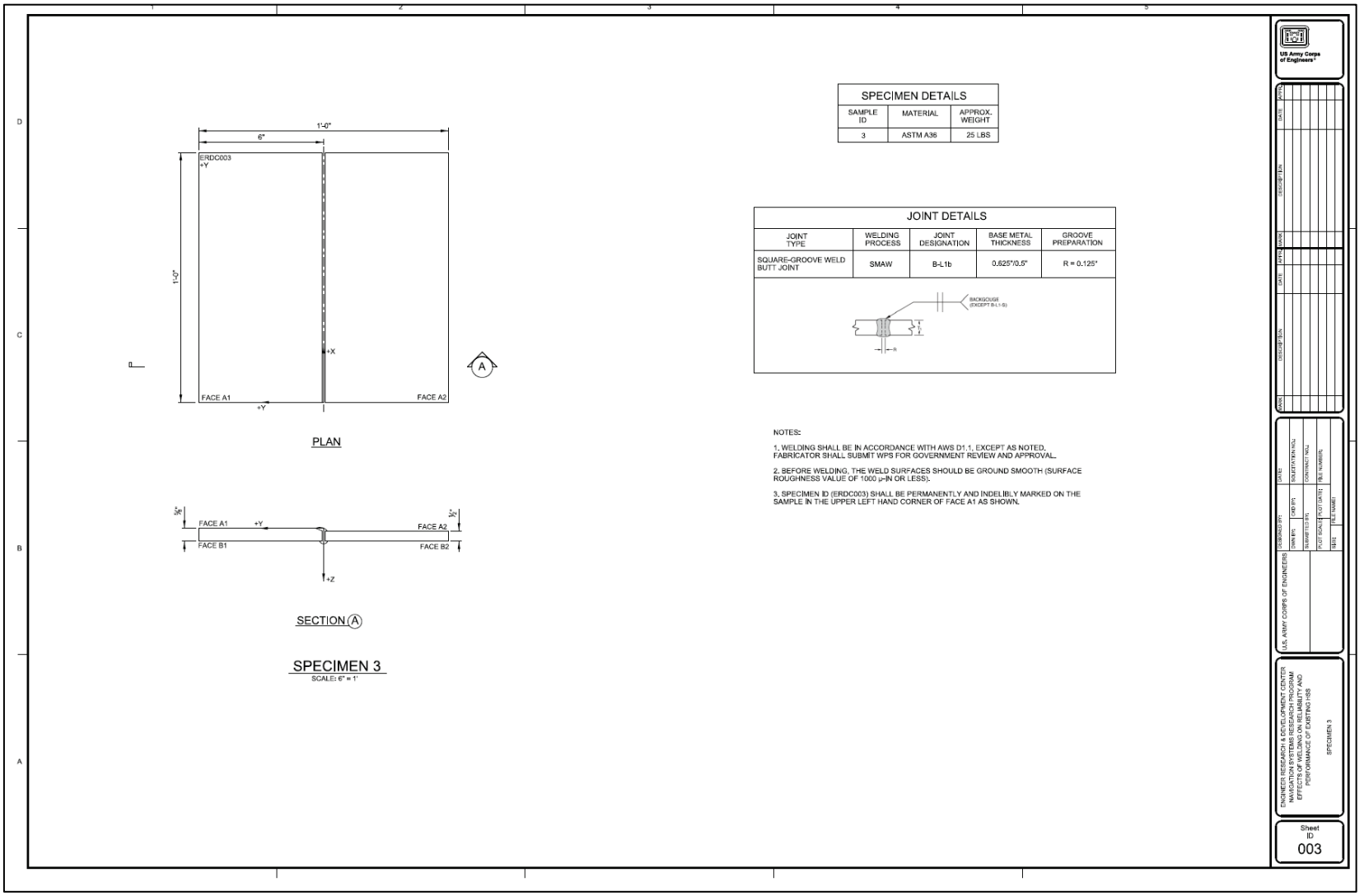


Figure B-6. Photograph of ERDC 003.

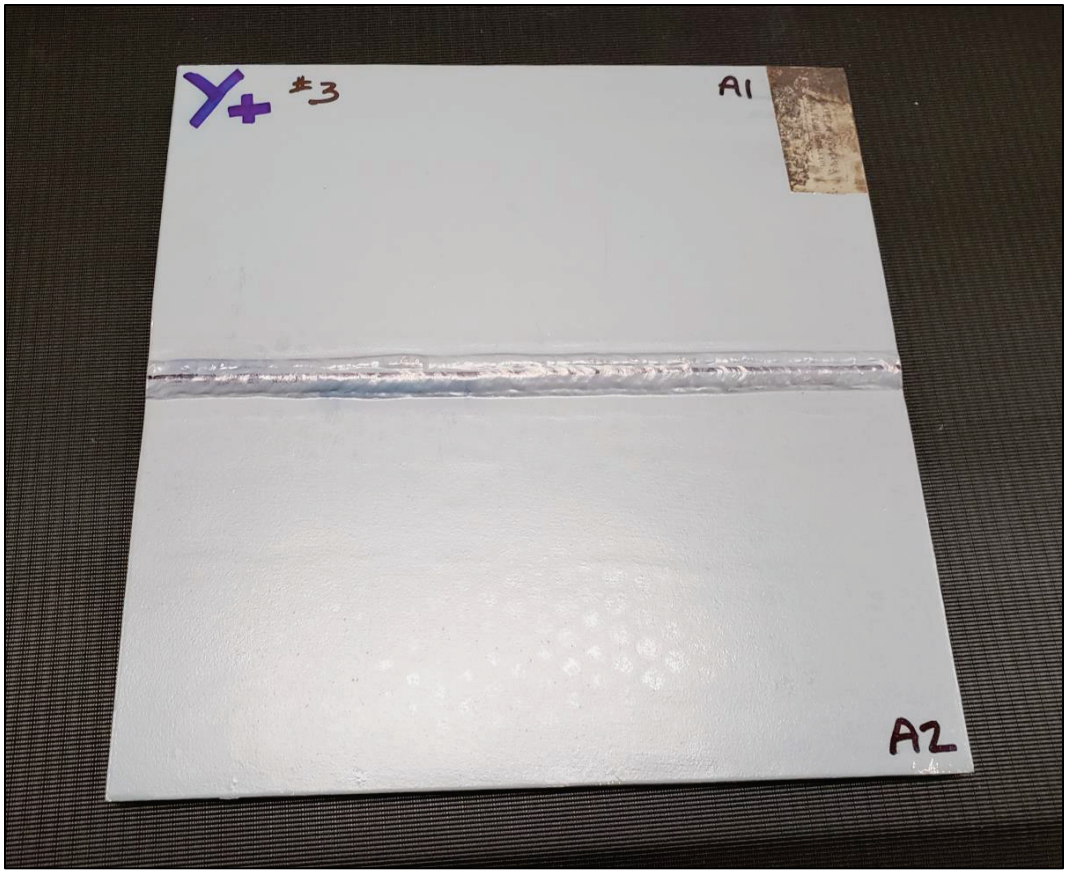


Figure B-7. Fabrication drawing for ERDC 004.

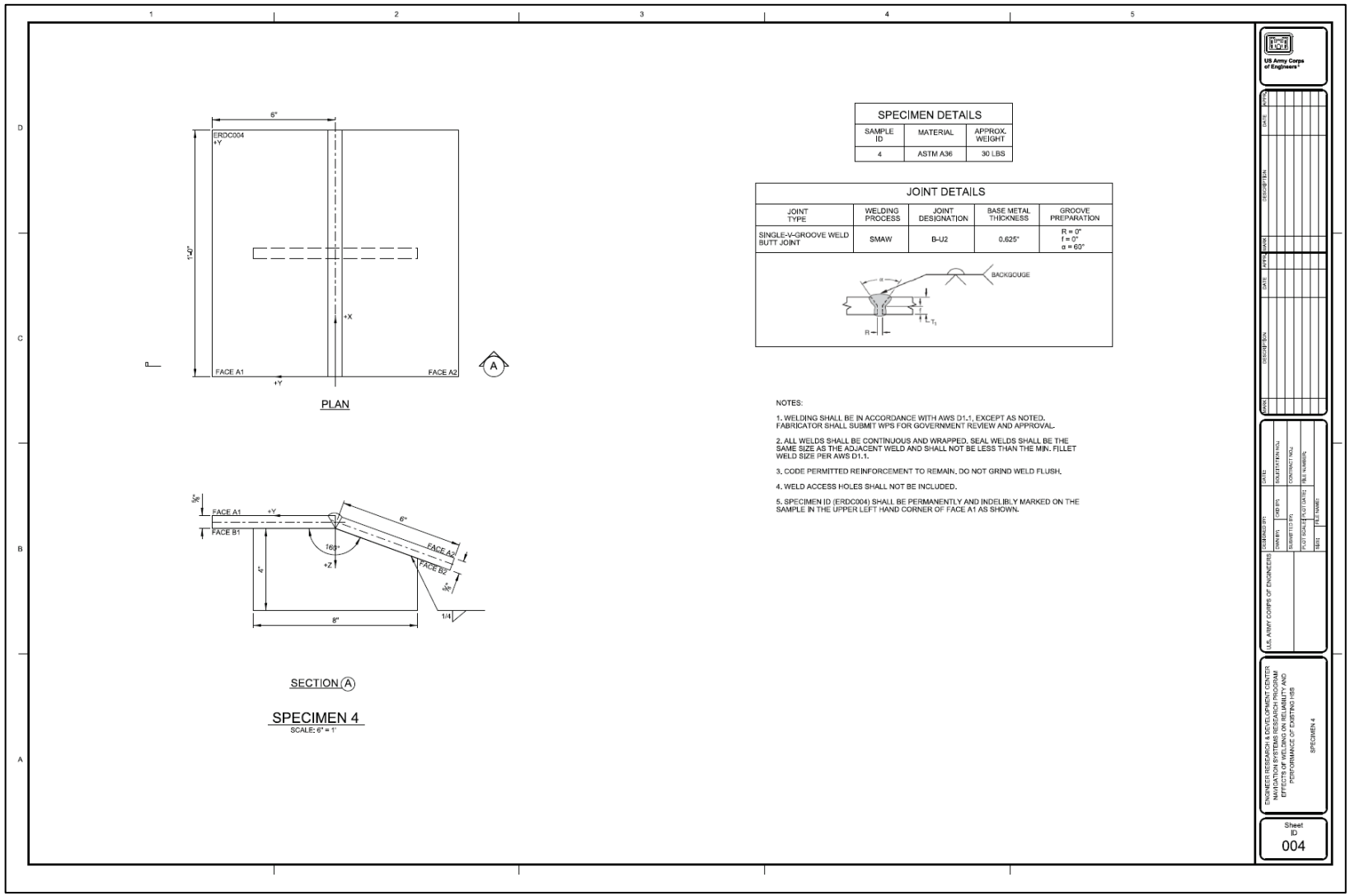


Figure B-8. Photograph of ERDC 004.

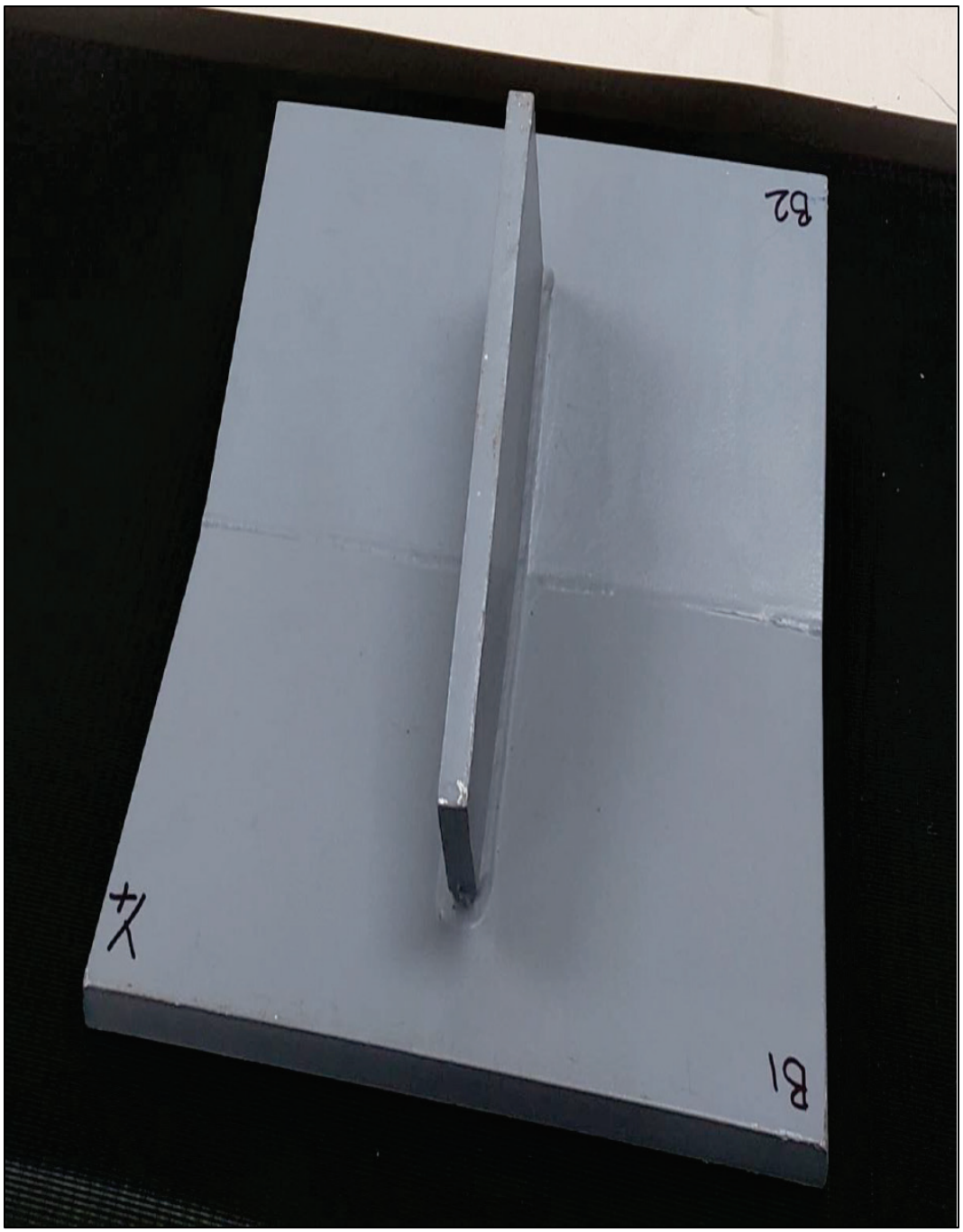


Figure B-9. Fabrication drawing of ERDC 005.

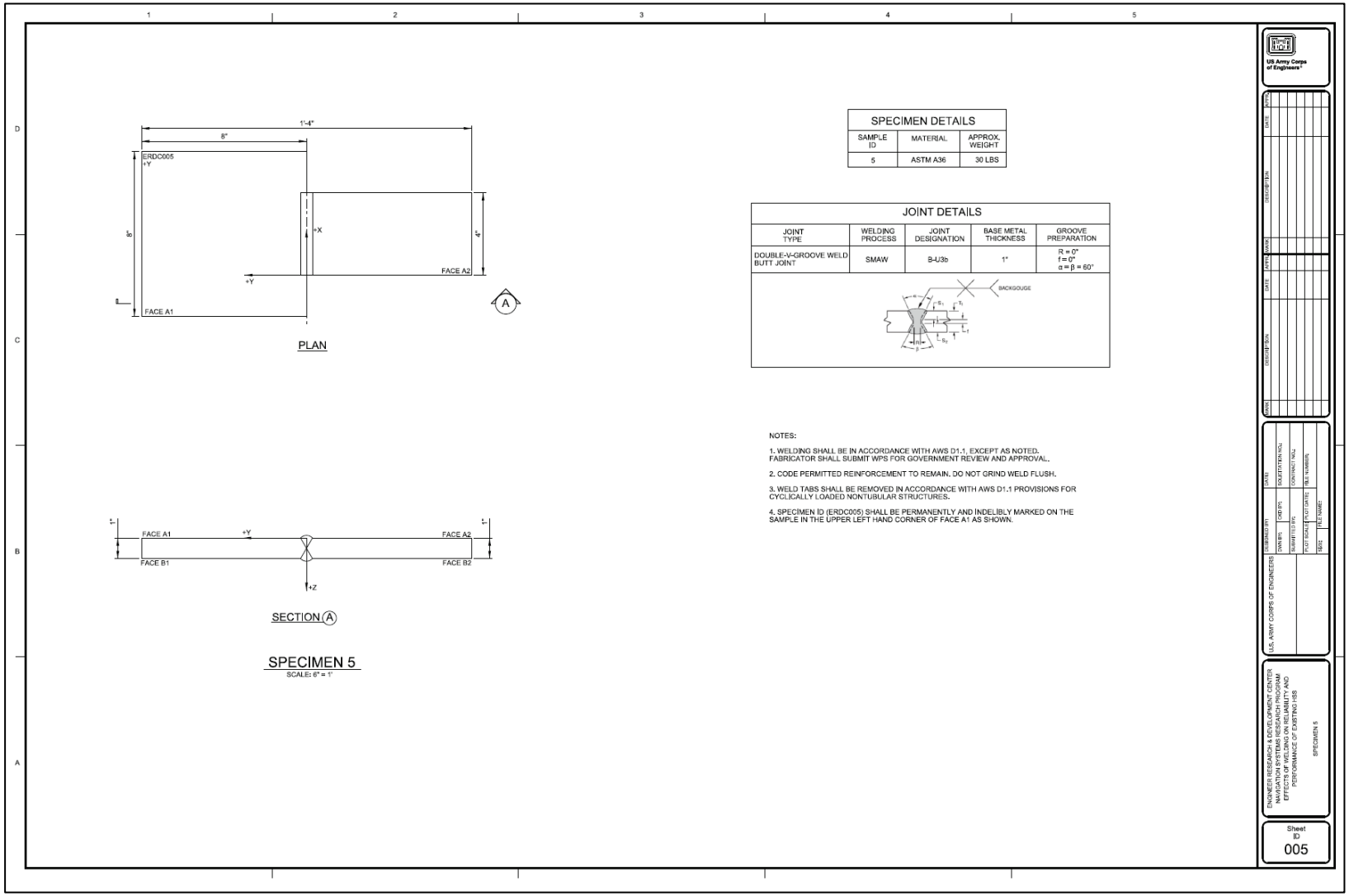


Figure B-10. Photograph of ERDC 005.

Figure B-11. Fabrication drawing of ERDC 006.

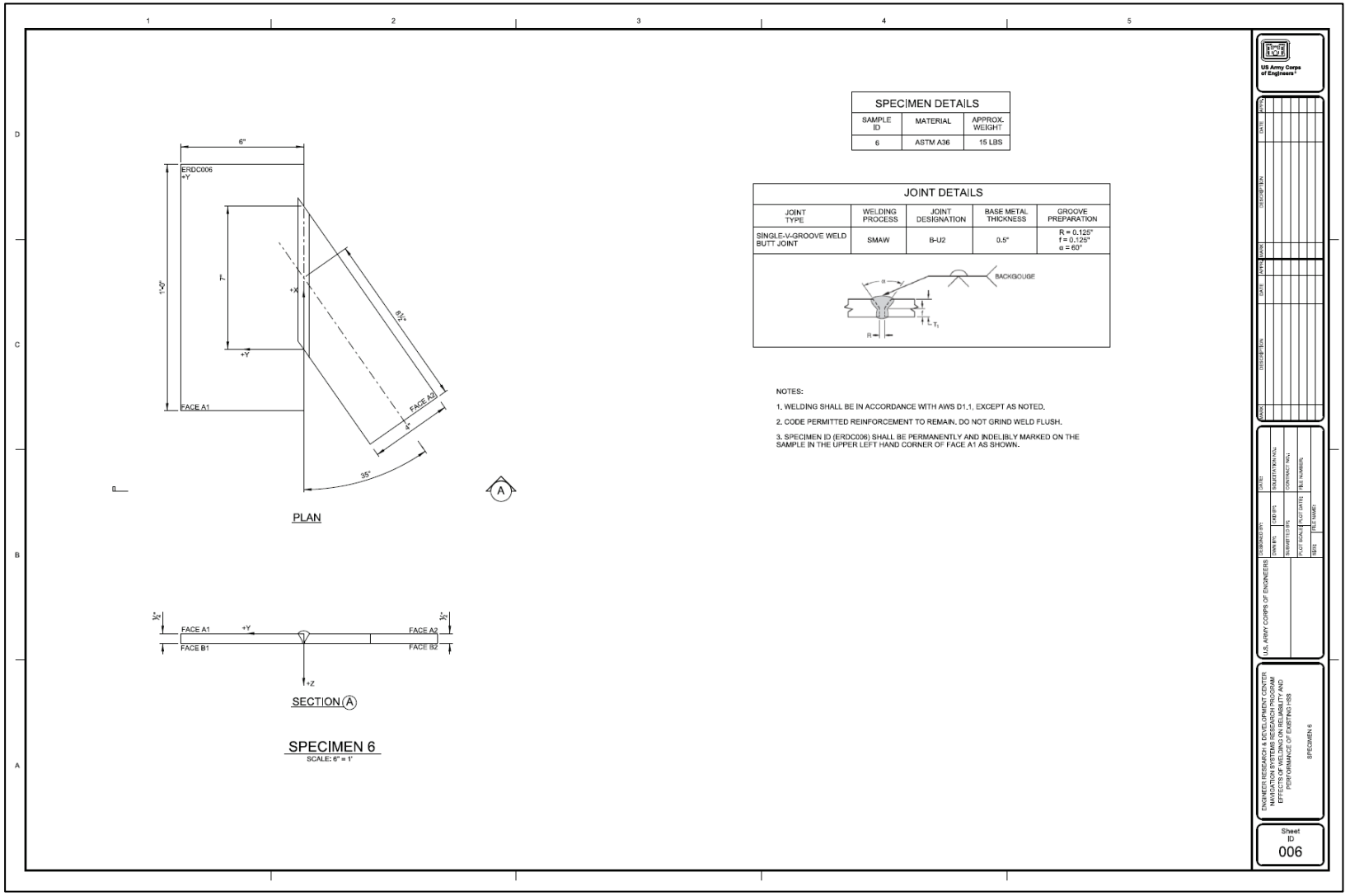


Figure B-12. Photograph of ERDC 006.

Figure B-13. Fabrication drawing of ERDC 007.

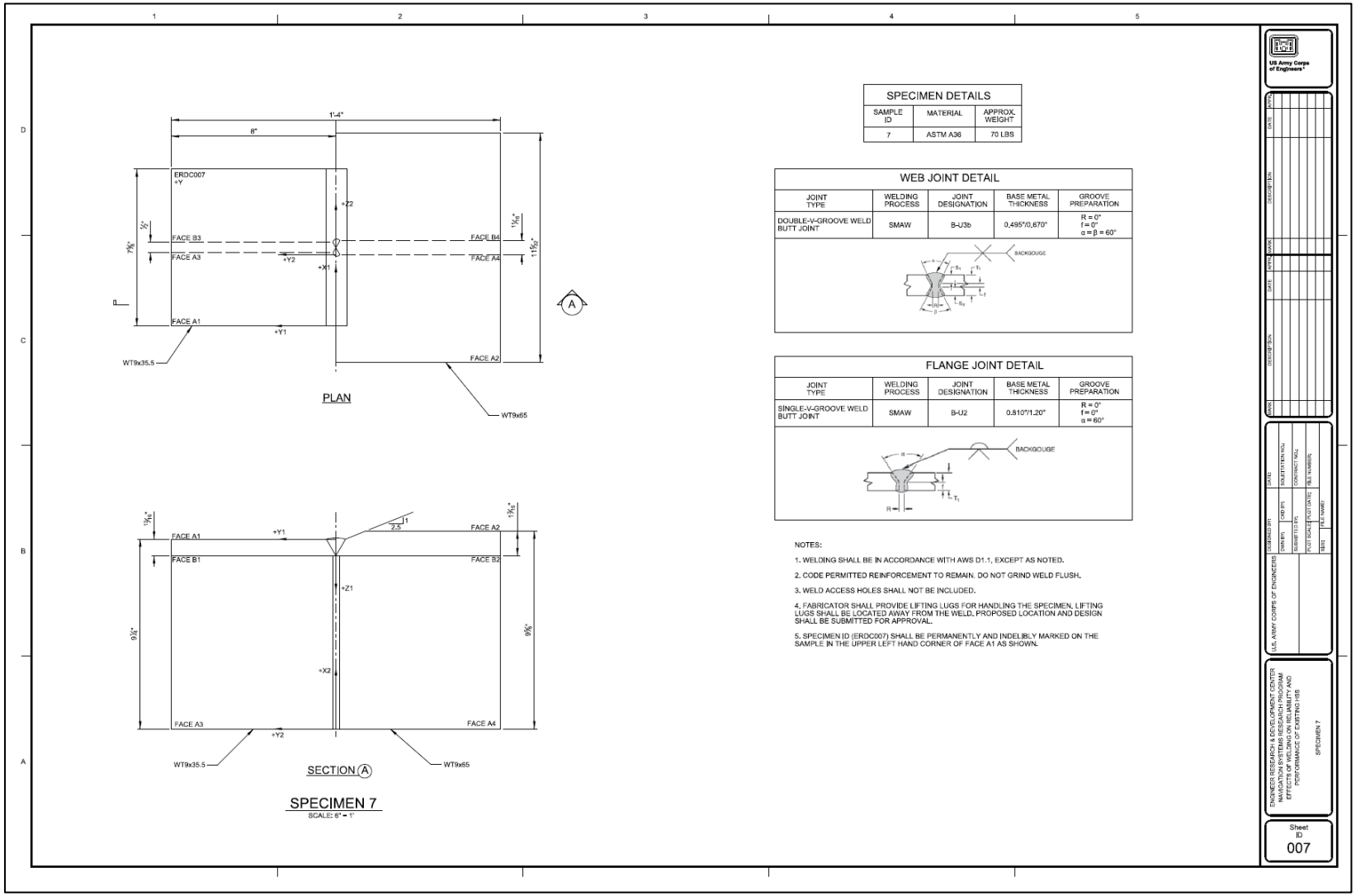


Figure B-14. Photograph of ERDC 007.

Figure B-15. Fabrication drawing for ERDC 008.

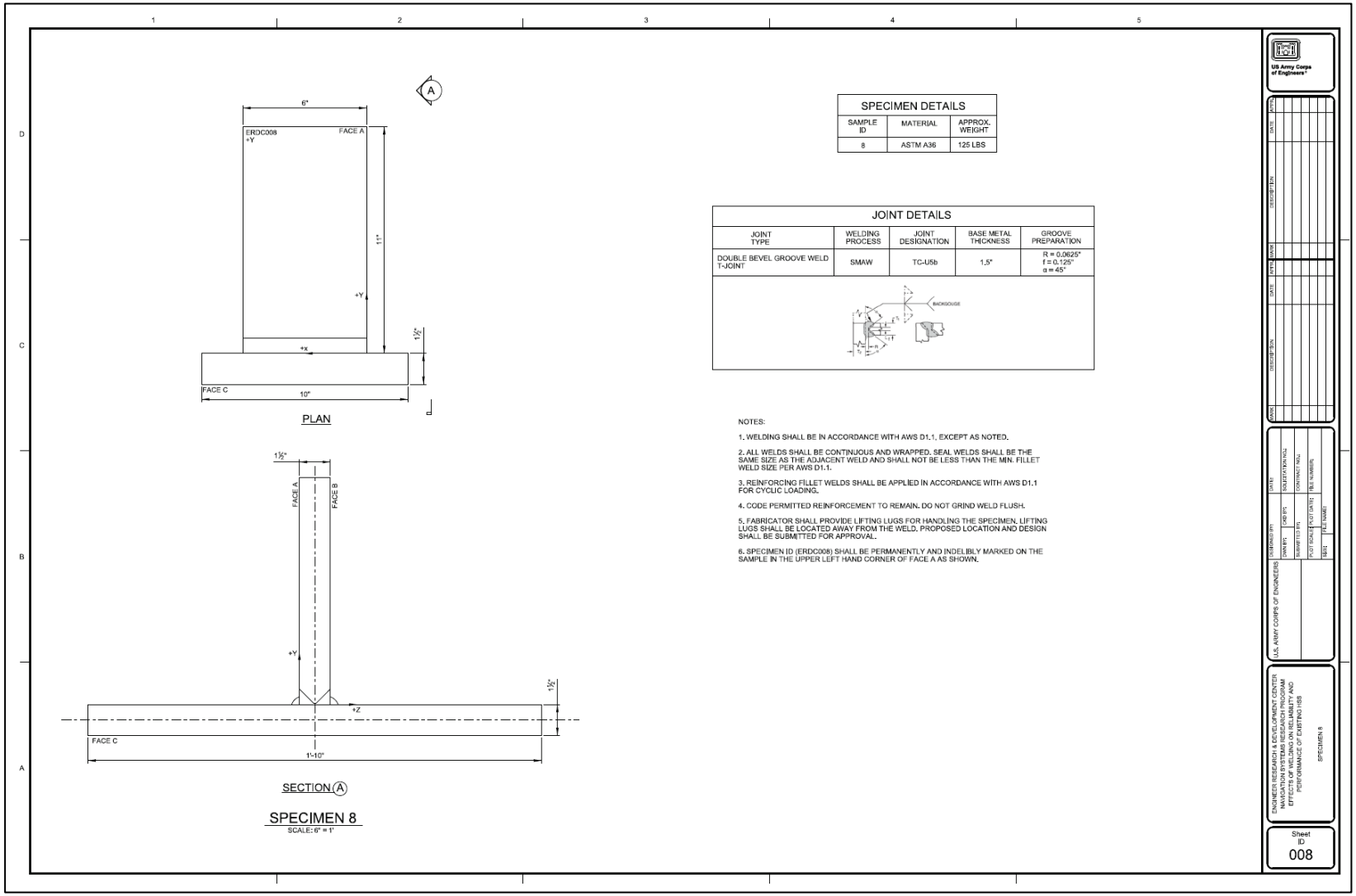


Figure B-16. Photograph of ERDC 008.

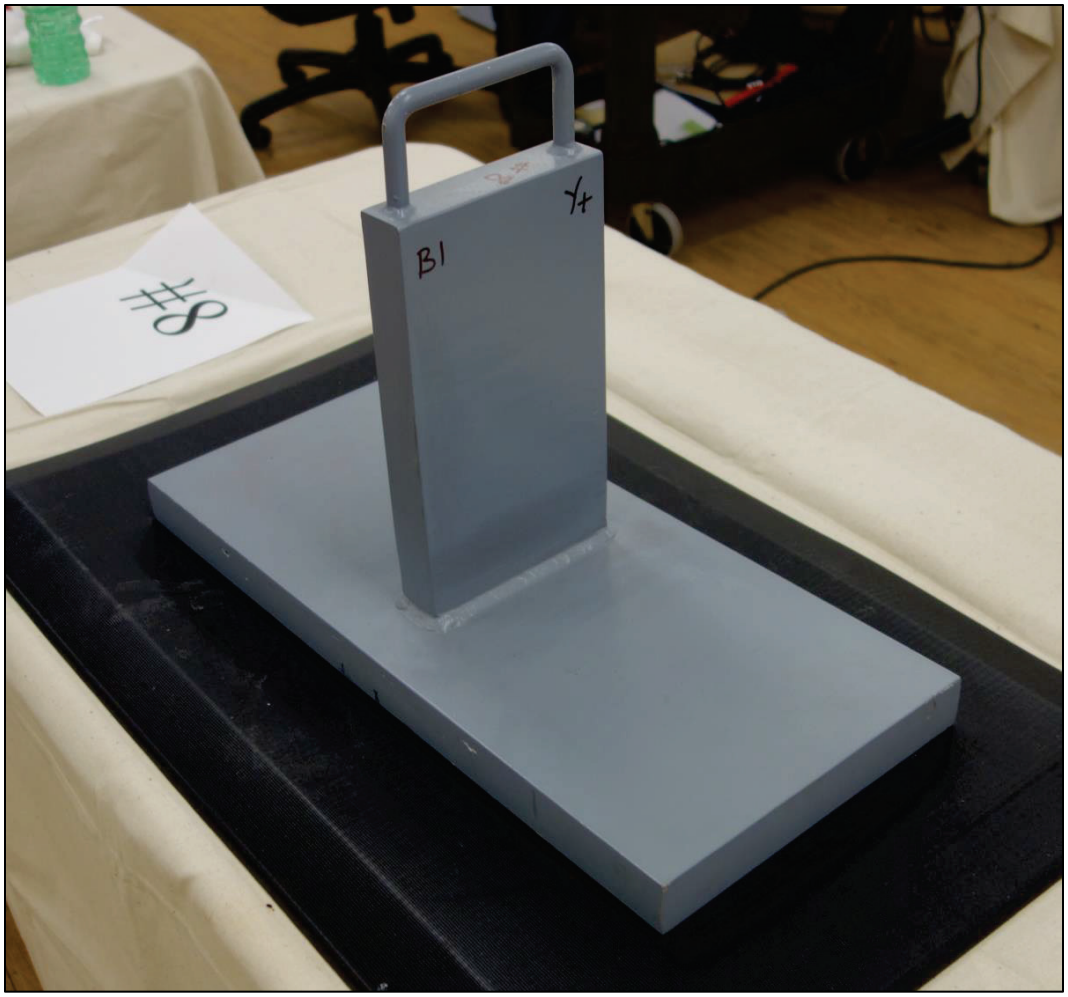


Figure B-17. Fabrication drawing of ERDC 009.

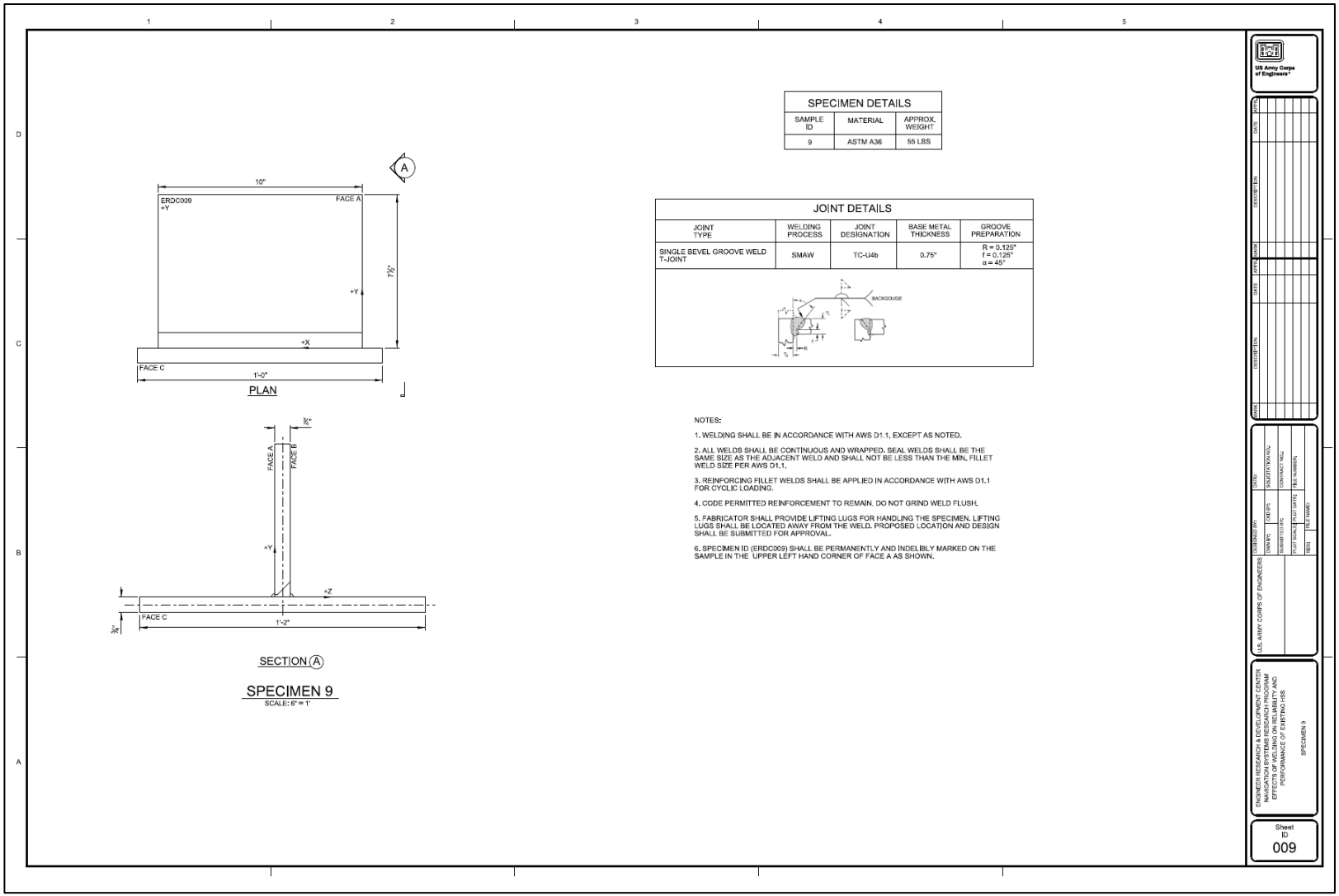


Figure B-18. Photograph of ERDC 009.

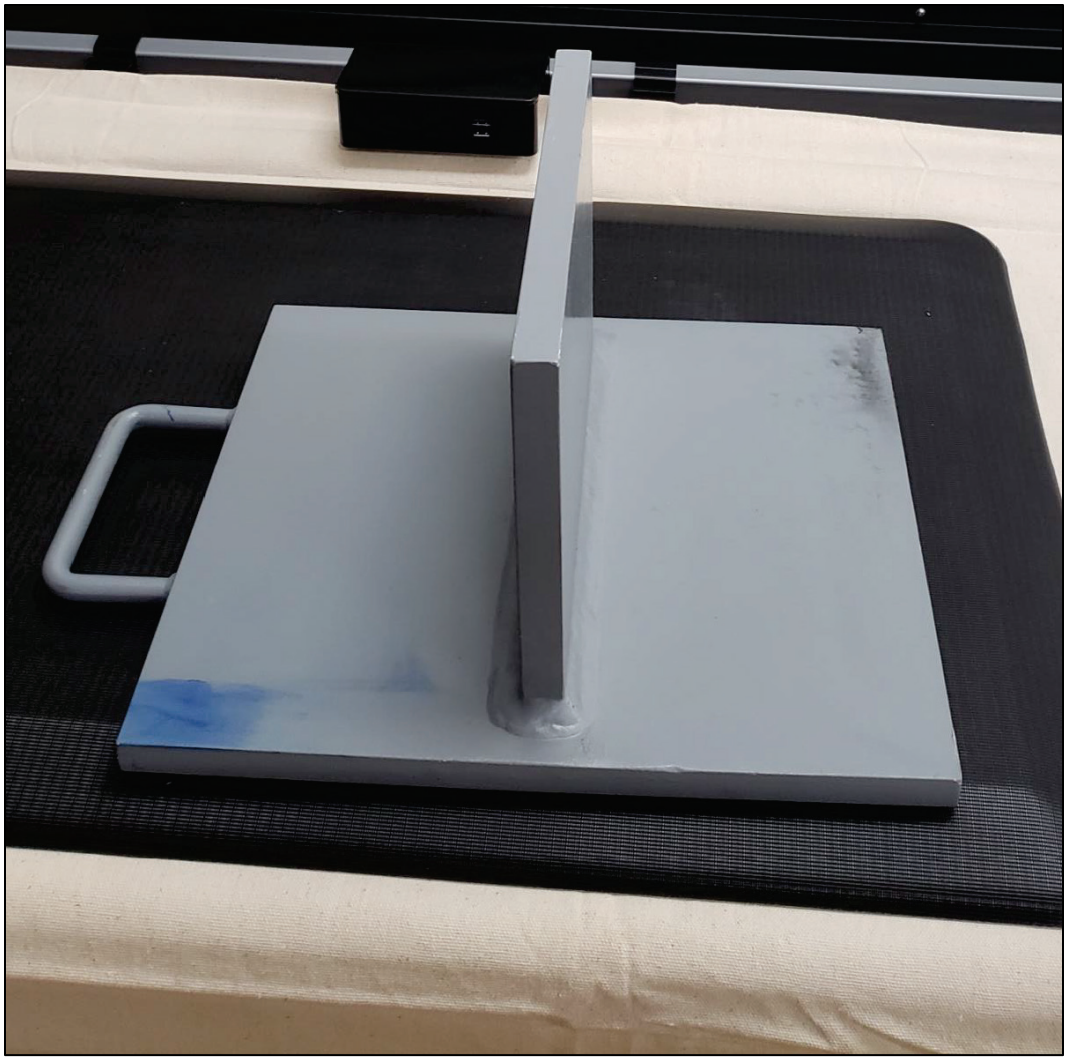


Figure B-19. Fabrication drawing of ERDC 010.

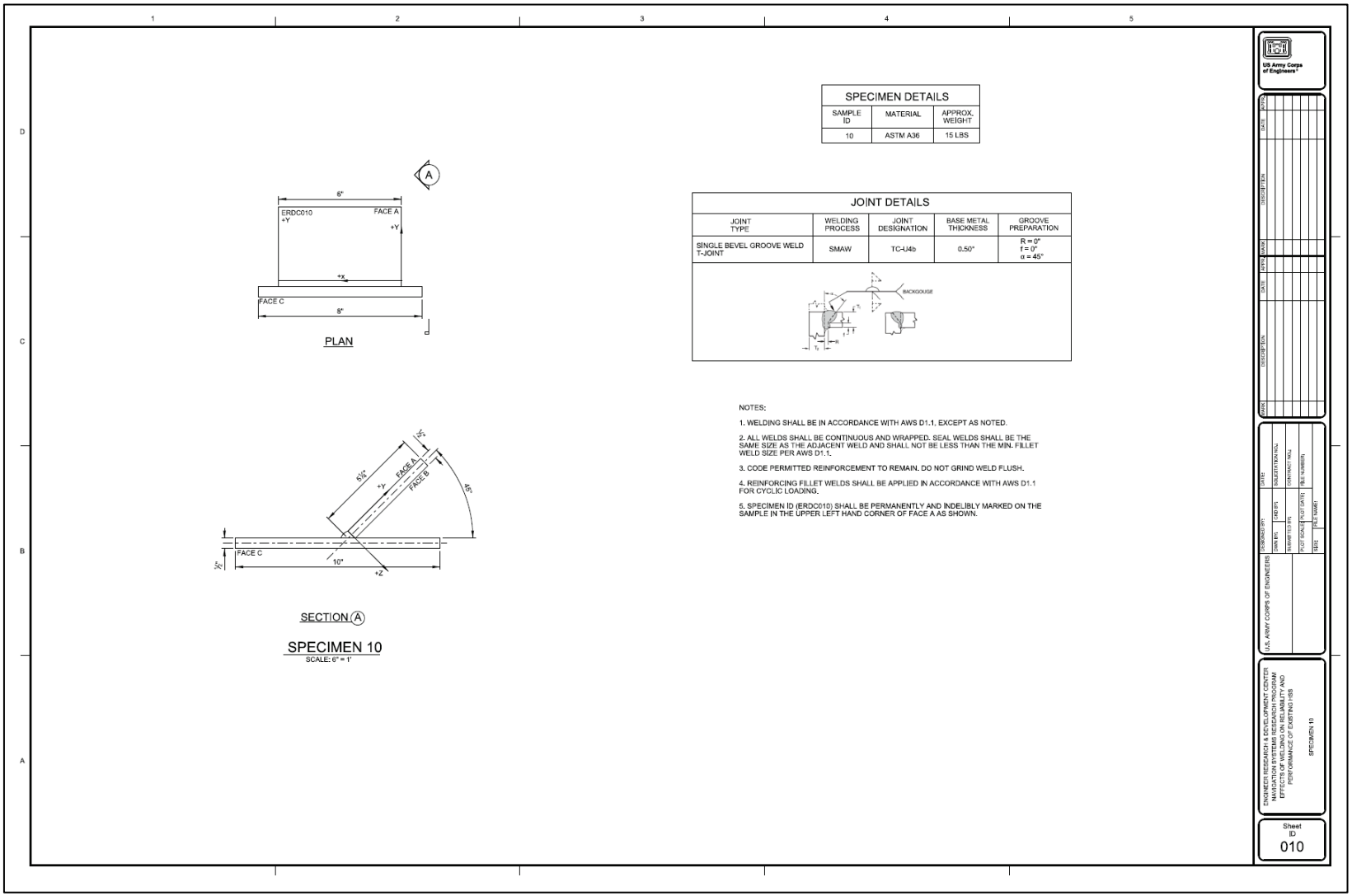


Figure B-20. Photograph of ERDC 010.

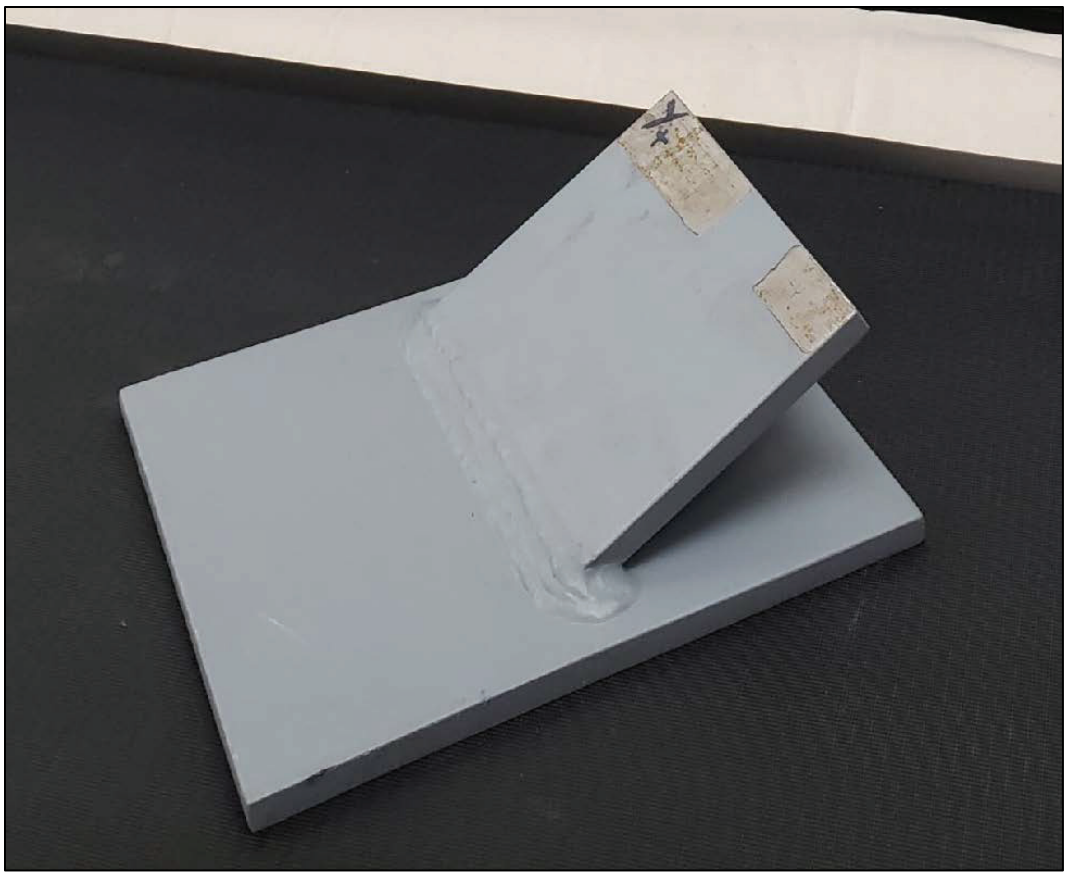


Figure B-21. Fabrication drawing of ERDC 011.

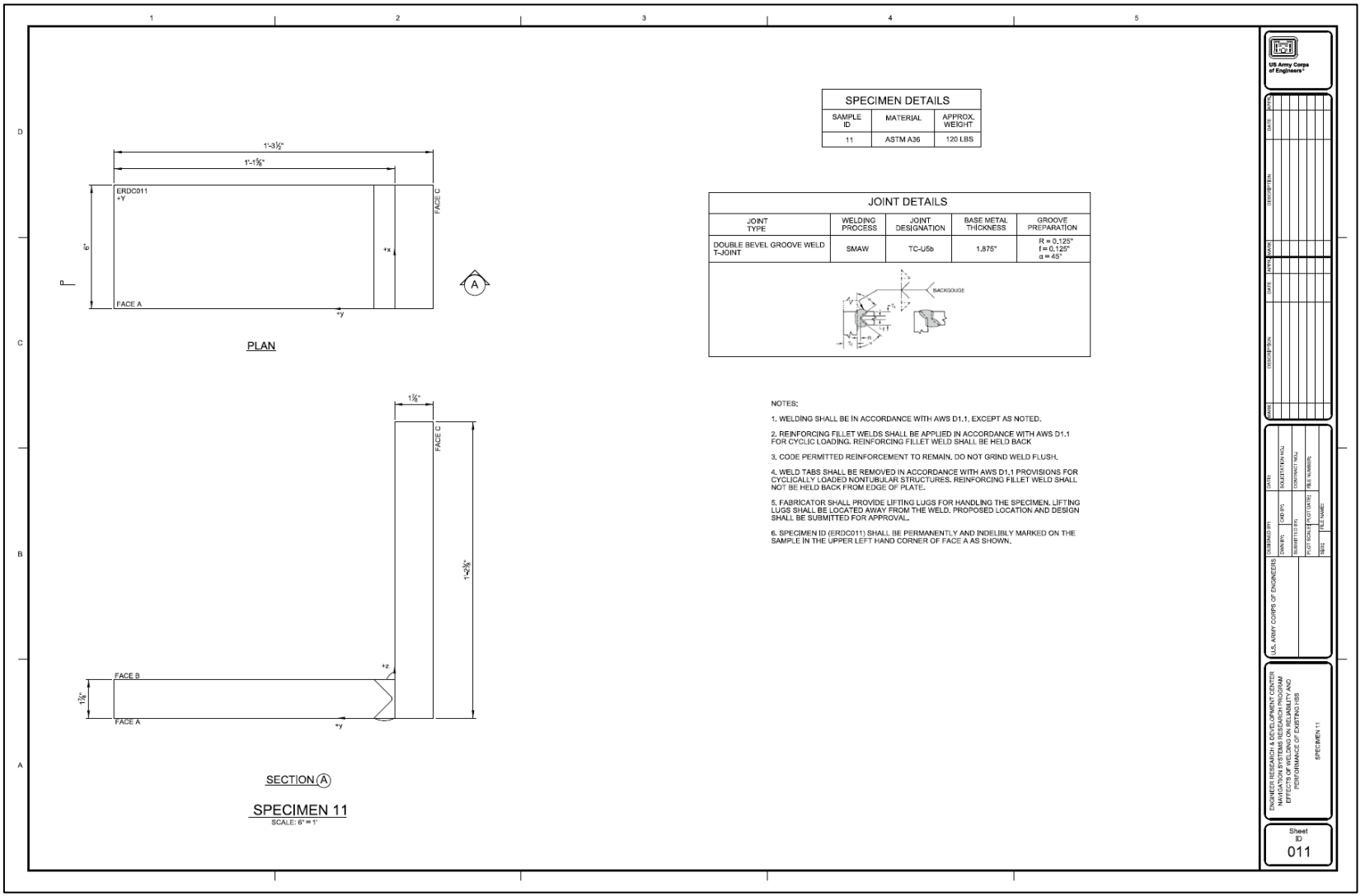


Figure B-22. Photograph of ERDC 011.

Figure B-23. Fabrication drawing of ERDC 012.

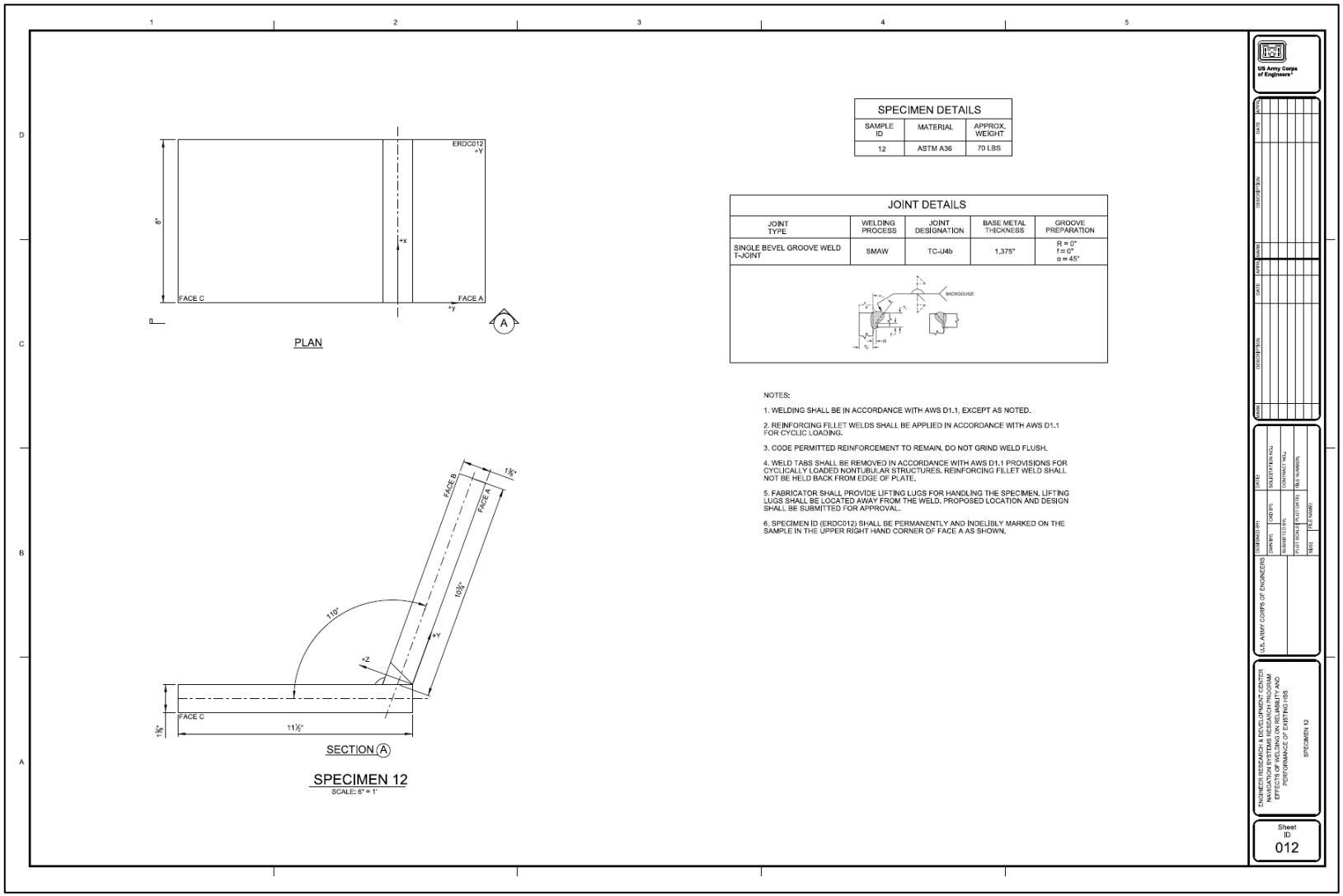
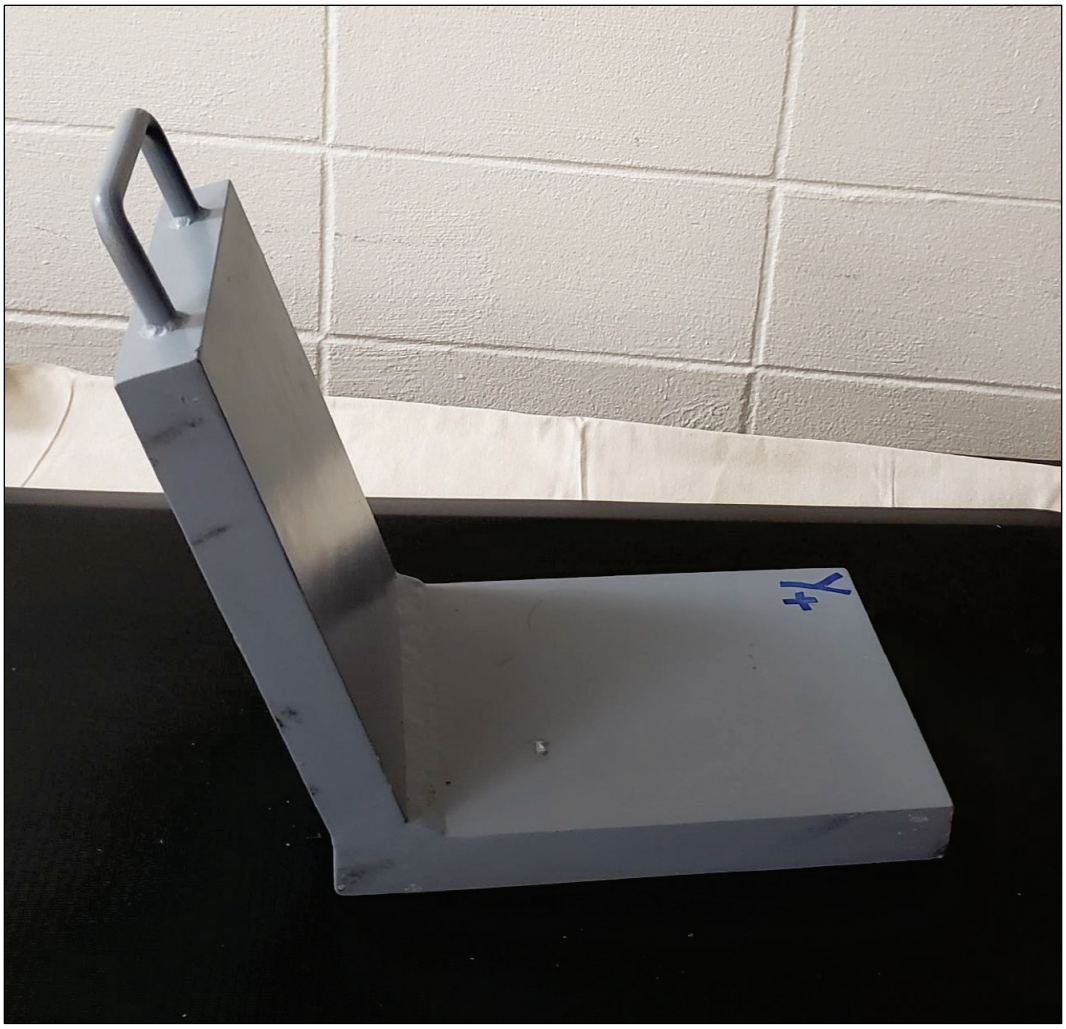


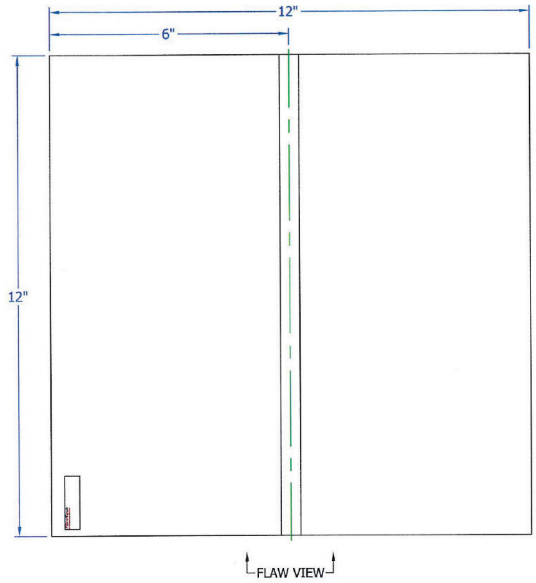
Figure B-24. Photograph of ERDC 012.



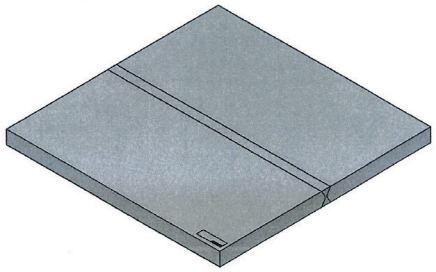
Appendix C: As-Built Drawings of American Welding Society (AWS) Specimens

This appendix contains the as-built drawings of American Welding Society (AWS) specimens. All of the drawings were supplied by FlawTech and are used with permission.

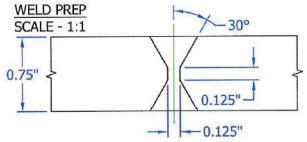
TOP VIEW
SCALE - 2:5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 1:1



[redacted]

[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

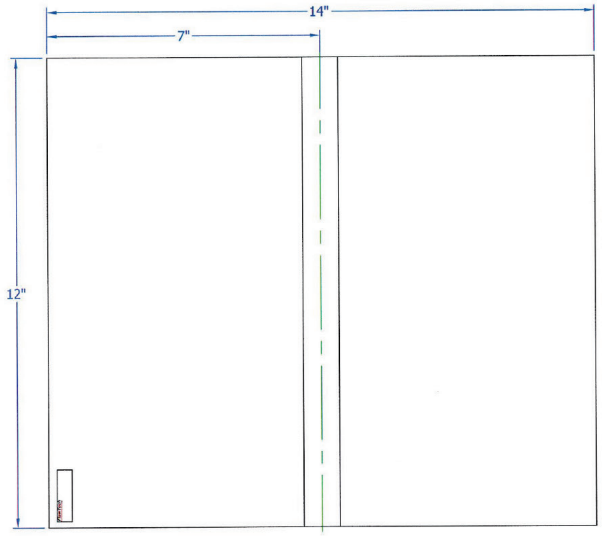
DESCRIPTION: 0.75"T X 12"L X 12"W PLATE SPLICE, DOUBLE V

INSPECTED BY: | DATE:

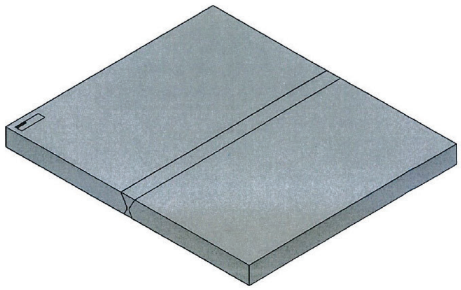
SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:

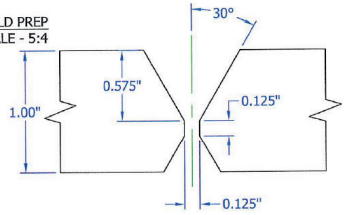
TOP VIEW
SCALE - 2:5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 5:4



[redacted]

[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

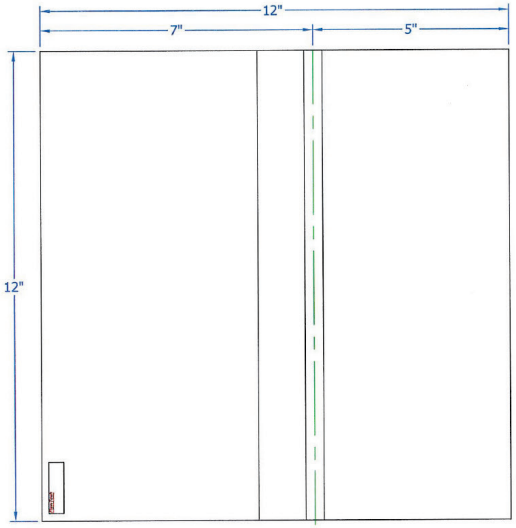
DESCRIPTION: 1"T X 12"WELD X 12"W PLATE SPLICE, 2/3 - 1/3 DOUBLE V

INSPECTED BY: | DATE:

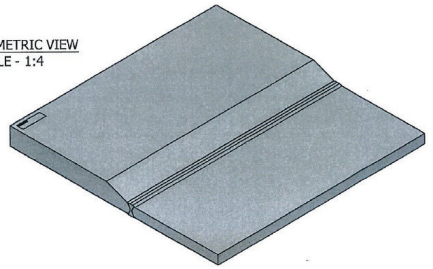
SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:

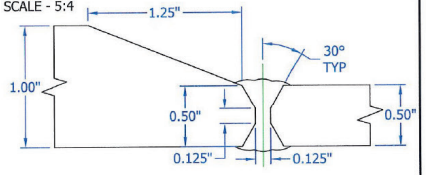
TOP VIEW
SCALE - 2:5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 5:4



[redacted]

[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

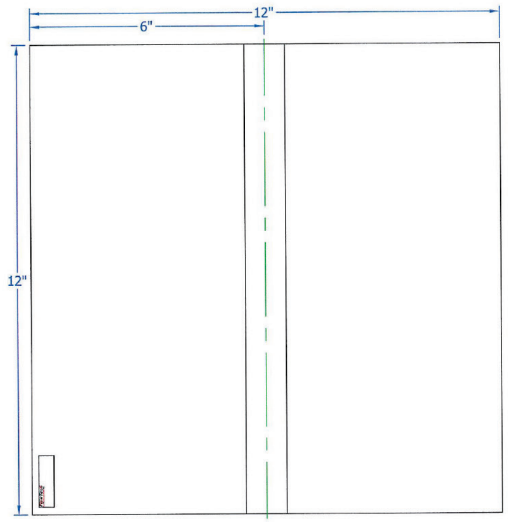
DESCRIPTION: 1.0"T TO 0.5"T X 12"L X 12"W PLATE SPLICE THICKNESS TRANSITION

INSPECTED BY: | DATE:

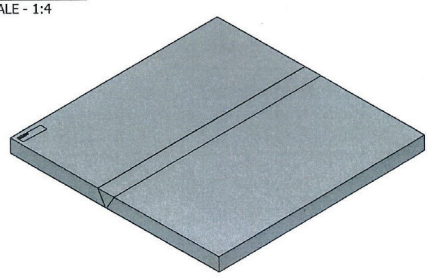
SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:

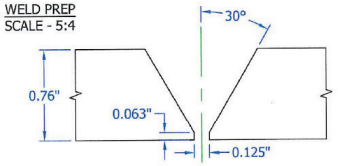
TOP VIEW
SCALE - 2:5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 5:4



[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

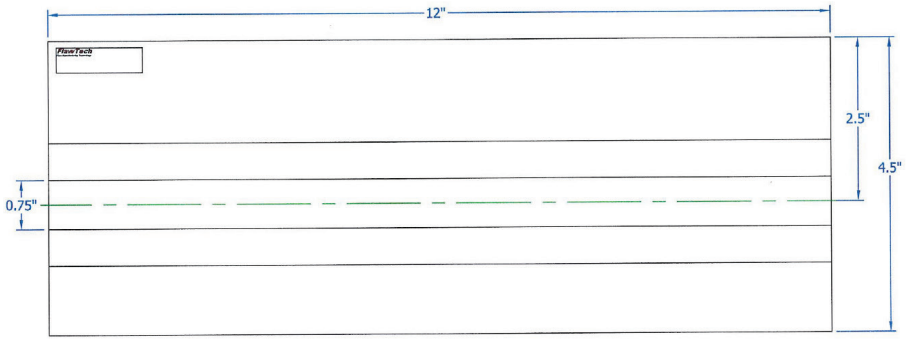
DESCRIPTION: 0.75"T X 12"L X 12"W PLATE SPLICE, SINGLE V

INSPECTED BY: DATE:

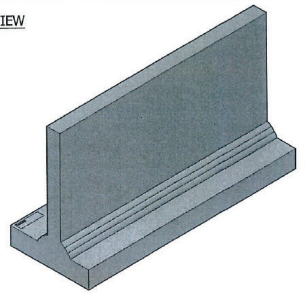
SCALE: SEE VIEW SHEET 1 OF 1

NOTES:

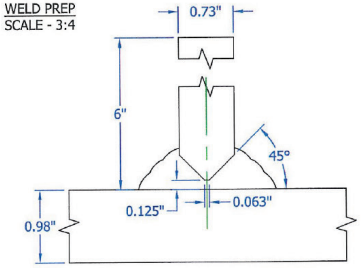
TOP VIEW
SCALE - 1:1.5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 3:4



[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

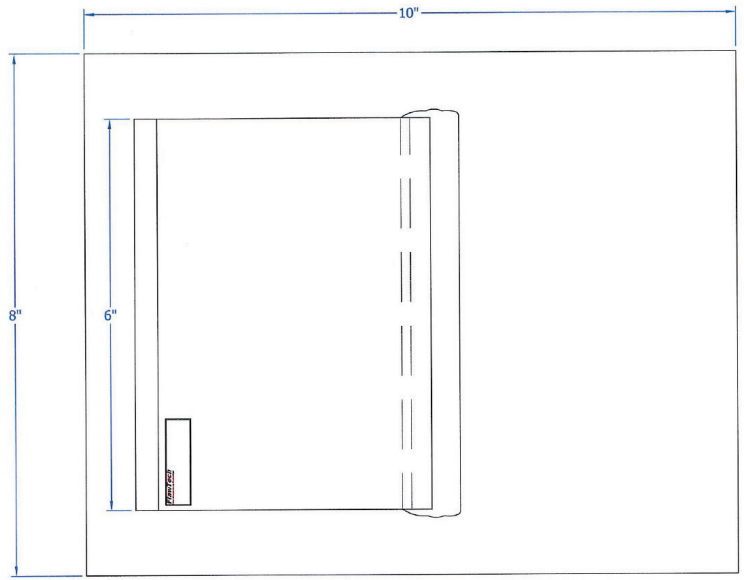
DESCRIPTION: 1.0" FLANGE TO 0.75"T X 6"H WEB X 12"WELD X 4.5"W PLATE

INSPECTED BY: | DATE:

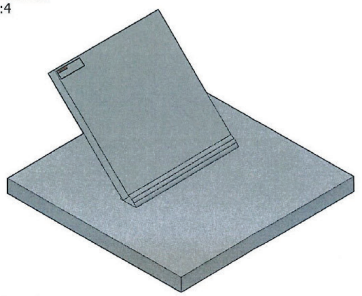
SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:

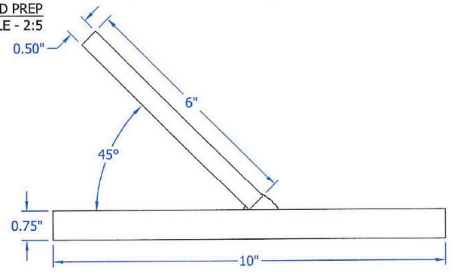
TOP VIEW
SCALE - 1:1.5



ISOMETRIC VIEW
SCALE - 1:4



WELD PREP
SCALE - 2:5



[redacted]

[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

TITLE: AWS D1.5 BRIDGE WELDING KIT

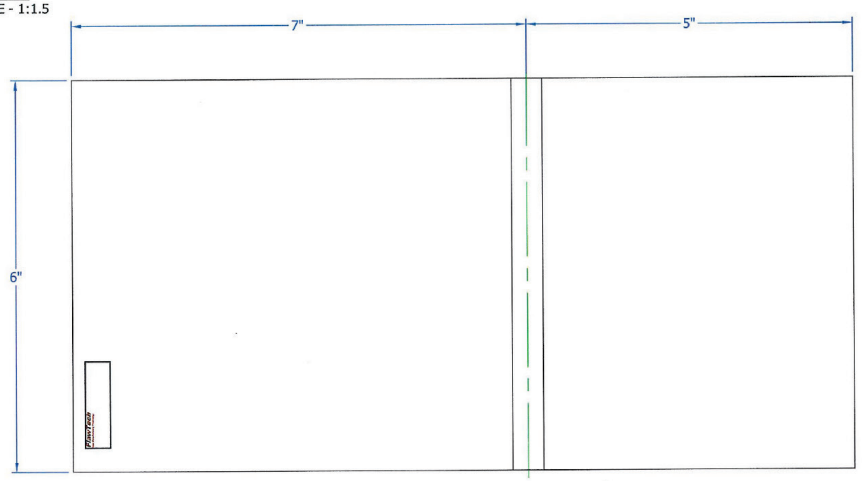
DESCRIPTION: 0.50"T CONNECTION PLATE TO 0.75" WEB

INSPECTED BY: | DATE:

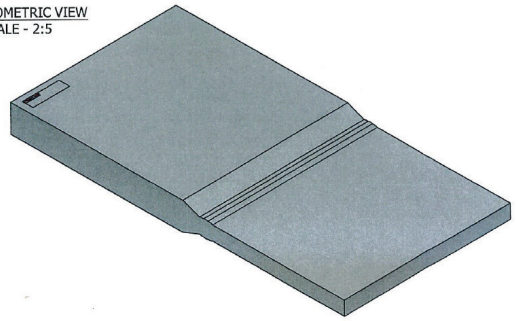
SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:

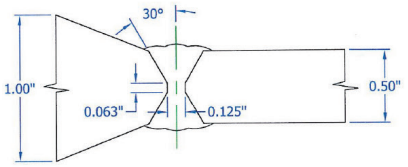
TOP VIEW
SCALE - 1:1.5



ISOMETRIC VIEW
SCALE - 2:5



WELD PREP
SCALE - 3:2



[redacted]

AWS-BK DRF

FLAWTECH
Flawed Specimen Manufacturing

SPECIMEN NUMBER:

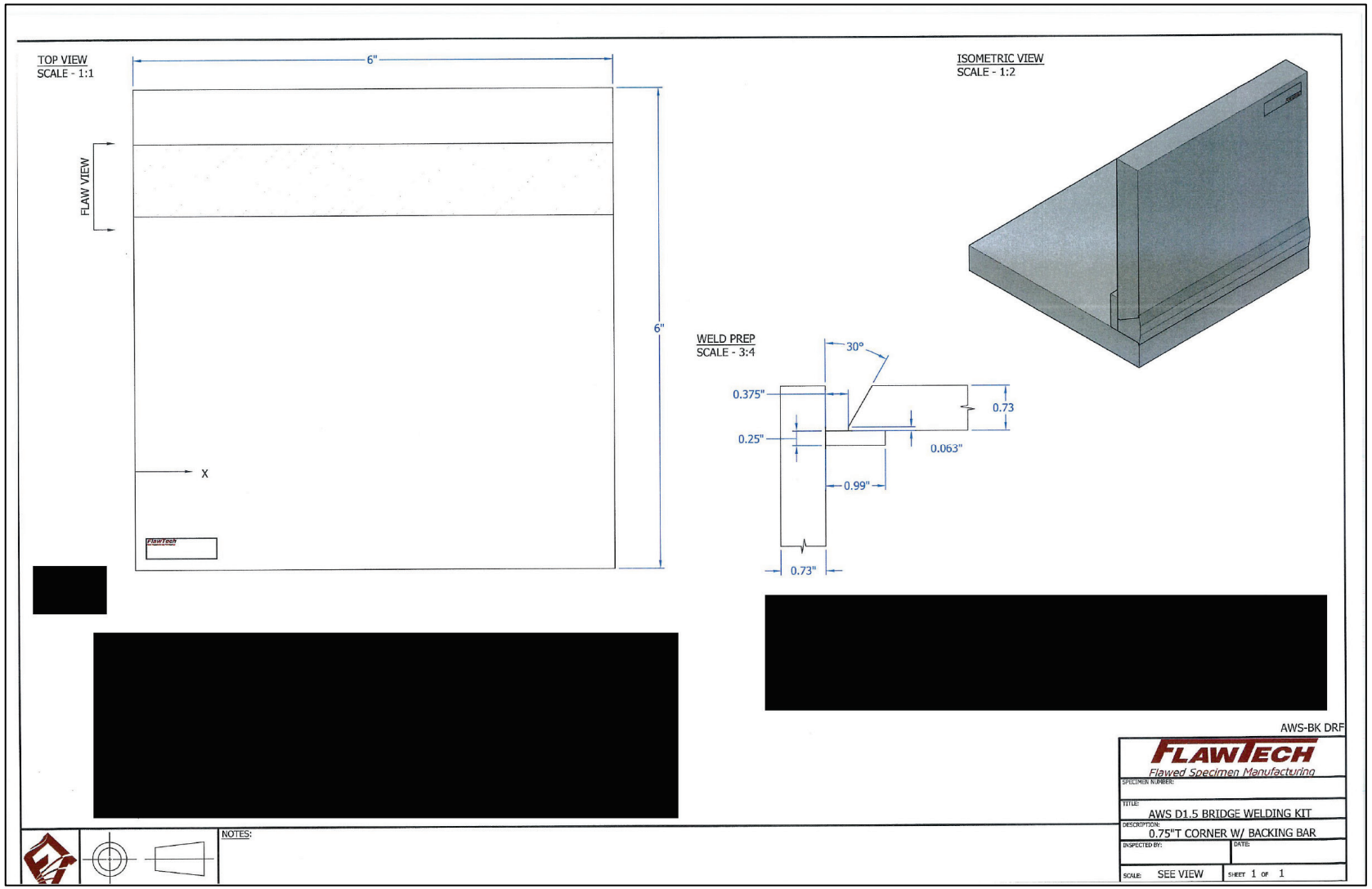
TITLE: AWS D1.5 BRIDGE WELDING KIT

DESCRIPTION: 1.0"T TO 0.50"T THICKNESS TRANSITION PLATE

INSPECTED BY: | DATE:

SCALE: SEE VIEW | SHEET 1 OF 1

NOTES:


TOP VIEW
SCALE - 1:1
6"
6"
FLAW VIEW
X
ISOMETRIC VIEW
SCALE - 1:2
WELD PREP
SCALE - 3:4
30°
0.375"
0.25"
0.73
0.063"
0.99"
0.73"
[redacted]
[redacted]
[redacted]
AWS-BK DRF
FLAWTECH
Flawed Specimen Manufacturing
SPECIMEN NUMBER:
TITLE:
AWS D1.5 BRIDGE WELDING KIT
DESCRIPTION:
0.75"T CORNER W/ BACKING BAR
INSPECTED BY:
DATE:
SCALE: SEE VIEW
SHEET 1 OF 1
NOTES:

Appendix D: Minutes of the Expert Workshop on Nondestructive Testing (NDT)

This appendix contains the minutes of the expert workshop on nondestructive testing (NDT).

Appendix 4: Minutes of the Expert Workshop on Non-Destructive Testing

USACE Weld NDE 2-Day Kickoff Meeting Notes (11/23 – 24, 2020)

Introduction 9am

Participants

- Frank Russo - MBI
- Greg Dunn - MBI
- Robert Connor - Purdue University
- Karl Frank - Austin Texas, Consultant
- Jordan Wind - NDT Consultant, Bureau Veritas
- James Kinnebrew - Mechanical Engineer ERDC
- Russell Kok - NDT Lab FHWA
- Ronnie Medlock – Subject Mater Expert Consultant to MBI
- Parrish Furr - NDT Consultant, Loenbro
- Phil Sauser - USACE
- Ramsay Bell - USACE
- Ray Momsen -Vice President Bureau Veritas
- Jason Ray - Electronics engineer EDRC
- Martin Schultz - USACE Engineer focused on Risk Analysis
- Tom Hay – NDT Consultant, TechKnowServ
- Hoda Azari - FHWA
- Leslie Campbell - Structural Engineer USACE, New Orleans

Frank Russo overview of Project

Task 1 - Expert Panel Review (11/23 to 11/24)
Discuss parameters leading to successful testing List essential parameters for testing of 12 joints

Task 2 - Develop Test Procedures
Develop Test procedures for 12 sample joints. Selected by USACE.

Task 3 - Conduct Testing
Test 12 joints at Vicksburg MS

Task 4 - Reporting

Integration with Parallel Project – MBI also has a parallel project with R Connor and SWRI to develop fitness-for-service (FFS) calculation guidance and examples

Jason Ray - Use of CIVA for Flaw Modeling Optimization
US Army Engineer Research and Development Center
Sensor Integration Branch

Presentation on what USACE is using NDE for currently
Guided Wave Ultrasonic Testing
PT Structures 40-80' long, embedded in conc in corrosive environments

Discussion on the use of CIVA for UT modeling of weldments. Jason is looking at:

- Using CIVA to optimize scan/probe to give best results for locating/sizing the flaws.
- Evaluate test plans determined by us using CIVA

- Discussion ensued about using CIVA to simulate different scan concepts, i.e. how many elements in PAUT to determine best procedures
- CIVA not good for complex structural connections (Ex. Jason ignores bottom plate for T-Joint in Specimen 1 for program feasibility)
- CIVA allows ease for varying/assigning parameters
- Jason is trying to develop a testing procedure in parallel with this task

DISCUSSION:

Martin - Trying to compare human testing vs. CIVA software, use workshop and actual testing to hone in on parameters and do further analysis for improve CIVA parameters. Certain parameters best optimized using software and some for trial and error

Karl - What does "best" mean? locating flaw? sizing flaw? Karl thinks height of flaw is most important. This is an opportunity to validate the CIVA program

Martin - defines best as maximizing probability of detection and minimizing error in flaw size

Tom - Agree w/Karl remarks, mirror the physical side w/modeling side for calibration. If modeling side mimics physical side, use modeling more exclusively for determining testing procedures (given confidence in modeling)

Group discussion about what we can / can't do in CIVA for FMCTFM? Discussion back and forth about varieties about wave sets to deploy in the specimen

Discussion by Rob Connor

- *Brought up NCHRP Report later shared in meeting*
- *Recommendations for D.1.5 for optimizing probe. Leslie sent it to Jason, but they haven't reviewed.*
- *Encouraged to reach out to Curtis Schroeder (experienced CIVA user).*
- *Sizing does not equal amplitude, important to keep in mind*
- *Will flaws represent tilt/Skew? Discussion was that there were no skewed or tilted flaws in the preliminary specimen design*
- *Schroeder Dissertation shared in meeting*

Ray

- *we (BV) use ES Beam Tool*
- *Ray doesn't believe CIVA can model some of the complex geometries. CIVA does simulations and EV does not.*
- *Reinforced point Rob made about skewed / tilted flaws and what comes from CIVA not really accurate compared to what happens when you put a probe on it.*
- *Russo commented - Ok to use Beam Tool, provides alternative solution/results for better calibration, improves "Scan Plan"*
- *Allows for complex geometries not feasible using CIVA*
- *CIVA does simulated testing where as ESBEAM does not*

Jordan

- *uses ESBeamTool a lot for developing techniques*
- *Familiar with CIVA (program to do many things)*
- *Used CIVA to determine transducers and wedge search unit projection*
- *Technique development - CIVA may cripple us rather than help us, limited user experience*

Russ

- *Considering optimizing techniques, flaw detection - there can be many different approaches*
- *Instead of picking random probe, most inspectors know which one to try first*
- *Suggest getting input from 4 expert inspectors as a starting point*
- *If modeled right, NDT right, you should get great results*

Jason - Agree, Flaw Location vs. Flaw Sizing, optimization for both of them individually. Make sure flaws are representative in real world. Reminder on CIVA limitations on complex geometry

Phil Sauser - Reliability of UT Methods on Hydraulic Steel Structure Welds (HSS Background)
USACE

Reviewed USACE role in civil works: Navigation, Flood Reduction, Hydropower, Recreation, Ecosystem Restoration

Hydraulic Steel Structures (HSS)
large inventory, variable type/size/use
harsh environments
Corrosion/section loss, impact/overload, fatigue, fracture

Discussion on Failures of HSS Structures and Samples

Gate Types
- Miter Gates - Typical, swinging gate, two hinges (top and Bottom), fatigue loaded as they are operated daily, high stress range/low cycles (design load is water load) Poor performing fatigue details don't last long, improper load paths that were not intended,
- Tainter Gates - Most common type of dam crest gate in terms of quantity, maintains water levels, radial pressure eliminates moment on gate, most members in compression, Vibration/Low Stress Range/High Stress Cycles,
- Lock Valves - Similar to tainter gate, uses to fill/empty lock, loaded from behind so everything in tension, difficult to inspect, uses up fatigue resistance quickly
- Lift Gate - Spillways and Locks
- Bulkhead/Stoplogs - Used for dewatering and emergency closures (not typically subject to fatigue loading),

Discussion on Engineering Regulation (ER) 1110-2-8157
- Welds not meeting AWS standard, need some type of evaluation. Conduct an Engineering Critical Assessment – focus is on life safety, economy, asset management. There is much variety in flaw size, type, stress conditions.
- Hard to standardize solution - different stresses, sizes, types, loading
- Combination of inspection methods (phased array, time of flight, etc.)
- PAUT Trials determine most are accurate, but plenty are undersizing and some are oversizing

Most HSS structures use vinyl coating (will be used for testing). Discussion on use of a SDH to determine how much gain to find flaws in the steel

Phil presented FFS (Fitness for Service) Variables (ASME 40 essential variables) - Current Research on Material Properties (toughness, stress intensity, flaw detection), look at variety of testing methods and CIVA support (Build less specimens, more modeling)

DISCUSSION:

Karl – SWRI developed the "backpack UT test" as a Go/No-Go inspection procedure after Point Pleasant collapse. An item studied was how to scan trough lead paint. We don't want to have to remove paint

Tom - Crank up gain until backwall reflection is consistent

Ray - Variation in coating thickness impacts attenuation. AWS has correction factors that can be used for paint influence. FFS technique doesn't require amplitude as such a big deal. Paint may not have as big an influence on sizing as it does for amplitude.

Tom - FFS amplitude is less important procedure, backwall 80% vs 60% likely won't influence sizing performance, does paint loss actually impact sizing? Doesn't think so

Parrish – Commented his experience is in-service assessments, some new construction/repair. Questioned if there will there be set of standards for in-service vs. new construction/repair

Phil - Not part of this task, but hopefully some of it can be applied to new construction/repairs

Martin Schultz - Estimating Strength and Toughness of Steel for Fitness for Service Analysis

- USACE procedure has been to repair all HSS with flaws / damage. Practice is costly.
- FFS can be used to define repair / no repair.
- BS7910 is identified as most appropriate method to conduct FFS for HSS
    - Three options for FFS: Option 1, Option 2, Option 3
        - Option 1 - Conservative and High Repair Cost
        - Option 2 - Middle ground
        - Option 3 - High Risk, less conservatism/repair costs
- Need material properties like yield, tensile, toughness, etc.
- Pre-1990 USACE has poor materials documentation
- Testing structures is destructive and may not be able to be used
- USACE lacks material info to do BS7910 analysis. Maybe can use approximate year of fabrication and characteristics to estimate material properties
- Developed Material Properties Database for HSS (Includes Bridges for additional data) 363 total observations

3 Types of estimates can be made for material properties

- Conservative (AASHTO guidelines for nominal steel strength of bridge, acceptable but does not address cost)
- central tendency - could overestimate strength/toughness leading to poor decisions
- probability distribution - calc probability of FFS, distributions condition on other characteristics

Bayesian Network approach has been looked at. This illustrates dependence among variables and range of values in database, key predictor variables (year of fabrication, specification, carbon equivalence, Rockwell hardness --> all nondestructive tests) not enough data to apply method really well, only 363 data points but need MUCH more

HSS - only get to look at once every 10-20 years

DISCUSSION:

Karl - FHWA Report can provide data source, NCHRP study with CVN Data. Need to consider crack growth analysis (after the meeting KHF provided some historical FHWA reports on materials. FMR sent to USACE)

Frank - Does not think the data is out there to support such a large database. Mentioned that some members are not as critical and suggests taking some samples from those locations to confirm material property assumptions

Martin - Thinks sampling is generally frowned on

Phil - Hard to get people to "cut up" their structure, decommissioning is a great opportunity. Martin's info could be used as a SCREENING tool

Rob - Need to determine your probability of the material properties, 50/50, 90/10? Data is needed to make good decisions. If staying conservative, FFS might not work effectively. When specifying CVN requirements those are a minimum (actual toughness likely higher), but it will be difficult to use more than minimum value unless willing to accept risk of using incorrect material properties

Martin - Welcomes other data from the Panel to help build the database

Ray - empirical data can't be "overstressed" for these structures, data does not constitute repairs and welding, highly variable for some components

Jordan - Recent project for railroad required material property determination for doing rehab/repairs

Panel collectively agrees to take samples from structure to determine material content

Phil - much easier to get chemistry than Charpy

Coffee Break 11:45 AM

Considerations for Use of FFS in Conjunction with NDT of CJP Welds
Robert Connor

- Analytical approach to evaluate if a given flaw will lead to failure under defined conditions, assumes flaws are "cracks or crack-like". BS7910, API 579, R6 are governing analysis documents
- FFS is powerful but there are many factors to consider
    - Defect position, joint geometry, stress range, yield and tensile, Kc (CVN and K correlations, Master Curve), defect aspect ratio, residual stress field
    - Crack growth / consequence of cracks growing is a consideration
    - Trick is interplay of flaw size and allowable stress but assuming crack-like flaws
- Preventing brittle fracture requires knowledge of
    - Flaw size
    - LAST
    - Toughness needed for weld and base metals
    - Loads needed - primary, transient, stress ranges
- Fatigue Crack Growth - Something stable today may not be acceptable tomorrow "Can I leave this FOREVER?"
- Residual Stresses - Can vary through component, BS7910 and API579-1/ASME FFS-1 gives guidance on flaw location
- Biggest Challenge for FFS - A lot of “knobs to turn” (thickness, stress ranges, a/c ratio, location, type of butt weld, etc.) Hard to determine a catchall scheme vs. looking at each structure individually
- Conservatism - cumulative effects that may only allow small flaws or indicate structure has failed
- FFS needs to consider inspection capabilities (reliability of Inspectors), NCHRP 14-35 less than 50% of actual racks reported correctly
- POD/Characterization is directly tied to the person performing the test, circumstances of test, component - etc.
- FFS methods need to successfully incorporate variability
- False positives an issue because they require repairs when not necessary
- NCHRP 14-35 aligns with other studies where tremendous variability (flaw characterization, flaw sizing, measured amplitude)
- FFS will not solve scatter issues which defeats the whole purpose

- Encourages putting inspectors in actual real-life scenarios to detect flaws (inside mockups of components) as it impacts flaw detection

Discussion:

Frank - While equipment is consistent, user is not. Need to try and eliminate user variability.

Phil - USACE interested in performance standards for inspectors

Rob - concerned with ability to accurately detect flaws consistently

Ray - Probability of detection is well-known established field in NDT, data is only as good as participants qualification/experience

Parrish - Suggests independent 3rd Party Performance Testing

Rob - Believes testing is not great enough for consistent accuracy

Ray - Disagrees, many states require certification for NDT

Karl - Scatter of NY inspectors similar to what Rob presented regarding variability

ASME Fracture Mechanics - PAUT
Parrish Furr

- Two different acceptance criteria: Workmanship Acceptance (Amplitude, Length, Type) and Fracture Mechanics Acceptance (Location, Size)
- Parrish developed a flowchart for technicians to determine inspection strategy
- 1st Step - Surface or subsurface
- 2nd Step - Determine Size
- Any FFS strategy depends greatly on accuracy of defect determination
- In-service approach uses workmanship acceptance

DISCUSSION:

Tom - Testing may be dependent on equipment, different inspectors use different equipment which may skew results

Ray - need to make procedures broad enough to accept variable equipment, or specify the required equipment for eliminate variability

Collaboration to some extent, is required.

Russell - Suggesting to pre-determine best approach before doing testing, recommends collaborating before starting

Ray - collaboration should occur before testing, not after

Phil - If we wait too long to collaborate, then won't have enough time in Vicksburg

Lunch 1:10 PM

Discussion of essential testing parameters/overall approach
Ronnie Medlock

- Started welding bridges in late 1950's
- UT criteria established late 1960's
- Alt UT criteria established in 1970's
- PAUT adopted 2015 for D1.5

- Methods of Inspection
  - RT
    - Pro (easy to evaluate actual test, long history of success)
    - Con Not good for planar defects, cumbersome and slow, no depth
  - UT
    - Good for planar, highly portable (preferred by High Steel Tech's)
    - Requires technician skill and written report
  - PAUT
    - Faster scanning, data recorded and readable
    - Slower evaluation
- Surface Methods of Inspection
- Visual Inspection - All welds (100%), well suited for all but tight cracks
- Magnetic Particle Inspection (MT) - 10% of main member fillet/PJP welds, "enhanced visual"
- Dye Penetrant - No requirement, good at tight crack, surface evaluation
- D1.5 criteria have long and credible history of ensuring superior performance in bridge welds
- UT provides greatest flexibility for in service welds
- PAUT offers enhanced UT
- RT is "clunky" but also good, not strong for cracks and poor for laminations
- Surface - visual usually sufficient, MT and PT as required

Tom Hay

- Reviewed Codes, Standards, Methods - AWS D1.1, D1.5 and ASME BPVC and experience testing USACE HSS
- Type of Testing - Visual, magnetic particle, liquid penetrant, ultrasonic shear wave, phased array shear way, eddy current, alternating current field measurement (ACFM), radiography, magnetic flux
- Case Study 1 - Spillway Gate
  - Environment can influence inspection results
  - Statement of Work - type of inspection method and access type, Code to inspect by, acceptability by what code, photograph all flaw locations, written report
- Case Study 2 - Butterfly Valve
  - Welding two dissimilar metals (Casting to Forged steel)
- Sizing Challenges
  - Lead and Vinyl Coating Systems - Getting soundwave in/out of specimen
  - Surface conditions
  - Access/obstacles to encoder scanning
  - Budgets and time constraints
- What is desired FSS inspection outcome for fracture critical analyses?
- In what planes/dimensions are weld defects, or any material defects, required to be sized?
- Depending on the structure and loading conditions are certain dimensions more critical than others?
- It would be useful for NDT DME's to understand, or tutored, on how the USACE uses the inspection outcomes

Specimen Discussion (combined notes from Day 1 and 2)
Frank Russo

- 37-page pdf document emailed to all prior to meeting

- What are we looking for?
  - Base Metal - Laminations
  - Welding - Incomplete root penetration, lack of fusion, porosity, slag inclusion
  - Cracking - longitudinal/CL cracking, transverse, toe crack, root crack

ACTION - Check if all methods are feasible for each specimen before proceeding

Uniformity of equipment - Bring whatever equipment you typically use, not limit to certain type

Request to supply CAD files to testing team – FILES PROVIDED

- We are not constrained to AWS probes. D1.5 is a good place to start for general guidance
- Use D1.5 for a guide on what we are finding and recording, or we will be there forever
- Do detection scans within the limits of D1.5 - switch to higher frequency as needed
- How about focused / unfocused? Do we do detection scan unfocused then size with focused? Tom Hay / Ray Momsen discussion
  - Parrish - D1.1 2020 for focusing is a possibility
  - Jordan - we should use D1.1 2020 for focusing
  - Tom Hay recommendation to scan unfocused to detect. Then switch to focused or higher frequency. Use D1.1 2020 for focusing
  - FMR - do we need to specify an exact probe / encoder? Parrish says no the language can be written to get what is needed for sizing
- Raster ends of all welds manual both sides both faces to get required coverage
- Tom Hay - how to deal with HAZ? Comment from Jordan / Parrish use AWS or ASME recommendations use lesser of 1" or "t" for definition of HAZ
- Karl - Sequence of welding and direction of tensile stress impacts defects
- Jordan - Not practical to scan all specimens from all sides and angles
- Russell - Scans from all angles/sides is most accurate, but most expensive/ least practical

Specimen 1 - T-Joint

- All three methods ok for inspection plus straight beam/manual for transverse
- Use straight beam and manual scanning for transverse
- No need for TOFD for transverse
- Definitely test face A unless run into a lamination. Won't see anything different from face B given size of weld
- For sizing, first leg has advantages
- If higher fidelity is needed for sizing, then we should write that in the procedure

Specimen 2 - CJP Thick Plates

- Several index locations to get full weld coverage. 2 or 3 index positions on each side of weld. supplemental sizing w/focusing
- Face A from both sides and supplemental scanning sizing with tighter focusing
- Jordan - might need a custom calibration block
  - Ray asked for discussion with ASME calibration block.
- Conventional TFM will be tougher at this thickness
- Parrish to sketch up location/size of holes --> Submit to Frank so fabricators of samples don't locate flaws at those locations
- Good scenario for TOFD, no complications
- Conventional TFM will be tough at this thickness

Specimen 3 - Butt Weld of unequal thicknesses

- Discussion - pretty simple Face A both sides supplemental sizing
- More than 1 index point, Parrish suggest 2 index points to catch the upper beveled surface

Appendix E: NDT Procedures

This appendix contains the NDT procedures.

CONTENTS

1. PURPOSE PAGE 2
2. SCOPE PAGE 2
3. REFERENCE & STANDARDS PAGE 2
4. PERSONNEL REQUIREMENTS PAGE 2
5. EQUIPMENT PAGE 2
6. CALIBRATION PAGE 4
7. DETECTION TESTING PROCEDURE PAGE 5
8. FLAW CHARACTERIZATION PAGE 7
9. LENGTH SIZING PAGE 10
10. HEIGHT SIZING PROCEDURE (dB Drop Method) PAGE 11
11. HEIGHT SIZING PROCEDURE (Diffraction Method) PAGE 12
12. FULL MATRIX CAPTURE (FMC) & TOTAL FOCUS MATRIX (TFM) PAGE 14
13. REPORTING PAGE 15

1. PURPOSE

1.1. The purpose of this procedure is to establish the minimum requirements for performing Phased Array (PAUT), Total Focusing Method (TFM) and Full Matrix Capture (FMC) ultrasonic flaw detection, characterization, and sizing.

1.2. The requirements of this procedure apply to candidates participating in Task 7 of Contract W912BU-18-D-0007.

1.3. Task 7 of Contract W912BU-18-D-0007 is intended to evaluate the reliability of ultrasonic testing methods performed on in-service hydraulic steel structures.

2. SCOPE

2.1. This procedure is for performing PAUT, TFM & FCM manual raster scanning of carbon steel groove welds and heat-affected zones in carbon steel materials ranging from 0.50 in. to 4.00 in. in thickness.

3. REFERENCES & STANDARDS

3.1. USACE Contract W912BU-18-D-0007
3.2. AWS D1.5 2020 Bridge Welding Code
3.3. AWS D1.1 2020 Structural Steel Welding Code
3.4. ASME Section V Nondestructive Examination
3.5. ASNT CP-189 2016 Standard for Qualification & Certification of NDT Personnel
3.6. ASTM E1316-17 Standard Terminology for Nondestructive Examinations
3.7. ASTM E2192 Standard Guide for Planar Flaw Height Sizing by Ultrasonics
3.8. ASTM E2700-14 Standard Practice for Contact Ultrasonic Testing of Welds Using PAUT
3.9. ASTM E2491-08 Standard Guide for Evaluating Performance Characteristics of Phased-Array Ultrasonic Testing Instruments and Systems

4. PERSONNEL REQUIREMENTS

4.1. All personnel performing ultrasonic testing in accordance with this procedure shall be certified to Level II in Phased Array Ultrasonic Testing and/or the Full Matrix Capture (FMC) and Total Focusing Method (TFM) methods

4.2. Level III personnel performing ultrasonic testing shall also be certified to Level II in Phased Array Ultrasonic Testing and/or the Full Matrix Capture (FMC) and Total Focusing Method (TFM) methods

5. EQUIPMENT

5.1. Any equipment make, model and style may be used as long selected equipment combinations meets the minimum requirements of this procedure.

5.2. ULTRASONIC INSTRUMENT

5.2.1. The phased array ultrasonic instrument shall be pulse-echo type suitable for use with transducers oscillating at frequencies between 1 and 15 megahertz.

5.2.2. The instruments should be equipped with a calibrated dB gain control stepped in increments of ≤ 2 dB.

5.2.3. The instrument shall be equipped with a minimum of 16 pulsers and 16 channels (16:16). A minimum of 16:64 is required for electronic scanning.

5.2.4. The instrument display shall be equipped with A-Scan, B-Scan, C-Scan and S-Scan display options.

5.3. SEARCH UNITS

5.3.1. STRAIGHT BEAM SEARCH UNITS

5.3.1.1. Straight beam search units shall be between 1 and 15 MHz inclusive and should be capable of producing compressional waves from -30° to +30°.

5.3.2. ANGLE BEAM SEARCH UNITS

5.3.2.1. Angle beam search units shall be a minimum 16 element linear array between 1 and 15 MHz inclusive and may produce shear and/or compressional waves. Refracted ultrasonic waves shall be between 40° and 70°.

5.4. EQUIPMENT QUALIFICATION

5.4.1. SYSTEM LINEARITY

5.4.1.1. Vertical and horizontal linearity shall be validated every three months in accordance with ASTM E2491-18

5.4.2. INTERNAL REFLECTORS

5.4.2.1. Internal reflectors for each PAUT search unit shall be validated at a maximum 40-hour intervals

5.4.3. RESOLUTION

5.4.3.1. Search unit and instrument combination resolution shall be verified and recorded prior to first use and after any change of wedge, transducer, or cable

5.4.4. PROBE OPERABILITY

5.4.4.1. An element operability check shall be performed by the PAUT operator prior to initial calibration and use to determine if dead or defective elements are present. No more than 10% of the elements may be dead in a given aperture, and no more than two adjacent elements may be dead within a given aperture. This check shall also be performed upon each 8-hour period of use. In addition, each element within a phased array probe shall be evaluated to check for comparable amplitude responses throughout the aperture. Each element shall be verified to be within 6 dB of the element yielding the highest amplitude response. If the amplitude of any of the elements within the probe yields responses outside the 6 dB requirement, the element shall be declared dead.

6. CALIBRATION

6.1. Any suitable couplant may be used. The same couplant shall be used for calibration and testing.

6.2. Screen range shall be sufficient to cover the entire range of depths for the thickness and angles being used.

6.3. Any design of calibration standard may be used but shall be made of cardon steel and use 0.063 in. side drilled holes for sensitivity. Velocity, wedge delay and sensitivity calibrations shall be made at the testing station immediately prior to testing.

6.4. Any design of sizing standard may be used but shall be made of carbon steel and may contain side drilled holes and/or notches which are typically placed at 20%, 40%, 60% & 80% of material thickness. Calibration sizing standards should be ± 10% of test specimen thickness

6.5. A correction factor of 10 dB shall be added to the primary reference level to compensate for coating attenuation loss.

6.6. Sector scans should be used as the primary scan to optimize coverage and shall be configured in angular sweep increments of no greater than 1°.

6.7. A minimum three-point TCG shall be established to cover the entire range of depths for the thickness and angles being used.

6.8. The standard sensitivity level (SSL) shall be established at 50% ±5% of full screen height using 0.063 in. side drilled holes. This dB value shall be recorded as the primary reference level.

6.9. The automatic reject level (ARL) shall be defined as 5 dB over SSL, which equals 89% FSH

6.10. The Disregard Level (DRL) shall be defined as 6 dB under SSL, which equals 25% FSH Please see Figure 1 below for an illustration of Automatic Reject Level (ARL), Standard Sensitivity Level (SSL) and Disregard Level (DRL)

FIGURE 1

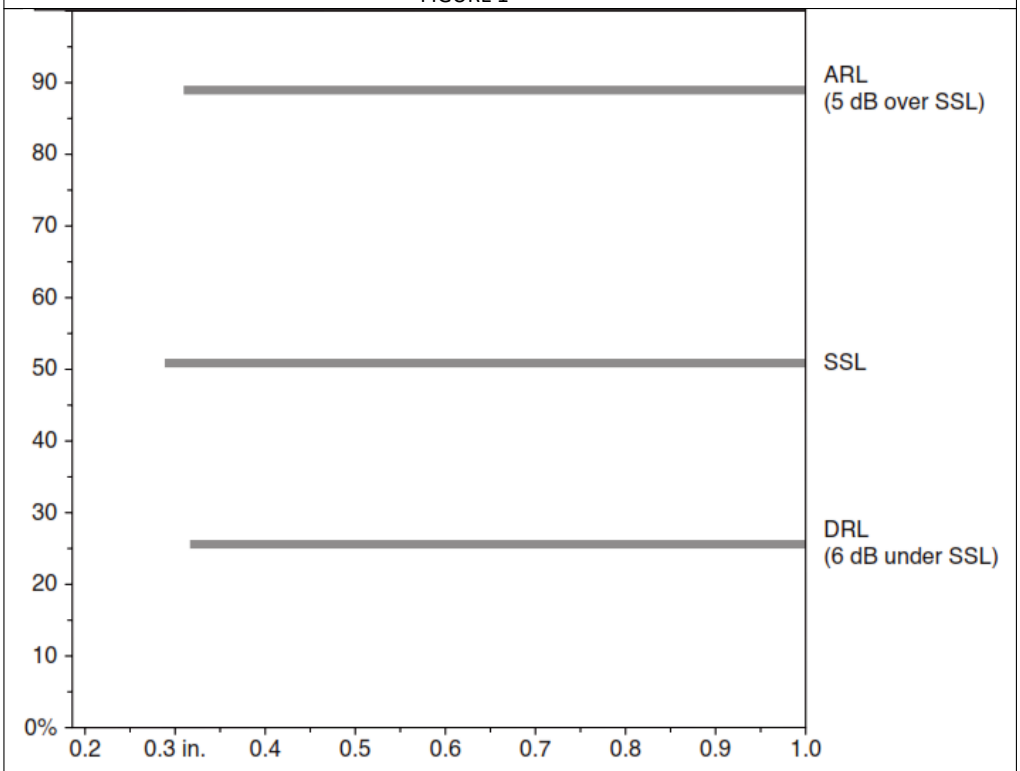


Illustration of DRL, SSL & ARL

7. DETECTION TESTING PROCEDURE

7.1. A "Y" accompanied with the specimen identification shall be clearly marked on each test specimen. The Y marking is used to identify the zero scan start and scanning direction. Please see Figure xxx for an illustration of the Y coordinates.

7.2. An "X" line for flaw location shall be marked on the test face of the weld in a direction parallel to the weld axis. The location distance perpendicular to the weld axis shall be the centerline of butt joint welds and the near face of the connecting member of corner and t-joint welds (the face opposite Face C). Please see Figure xxx for an illustration of the X coordinates.

7.2.1. STRAIGHT BEAM SCANNING

7.2.1.1. The entire base metal through which ultrasound must travel to test the weld shall be tested for laminations using straight beam compressional waves. If any area of base metal exhibits total loss of back reflection or an indication equal to or greater than the original back reflection height is located, the size, location and depth from face-a shall be reported.

7.2.1.2. Corner and t-joint welds shall be tested from face-c targeting the weld and HAZ area using straight beam compressional waves.

7.2.2. ANGLE BEAM SCANNING

7.2.2.1. It is the intention of this procedure that all welds be tested by passing ultrasound through the entire volume of weld and heat affected zones in two crossing directions. At a minimum, all welds shall be tested using a manual raster scan utilizing the patterns in Figure 2 below. Butt joint welds shall be tested from each side of the weld axis from face-a and face-b. Corner and t-joint welds shall be tested from face-a, face-b and face-c. Please see Attachment 1 for an illustration of face-a, face-b and face-c.

FIGURE 2

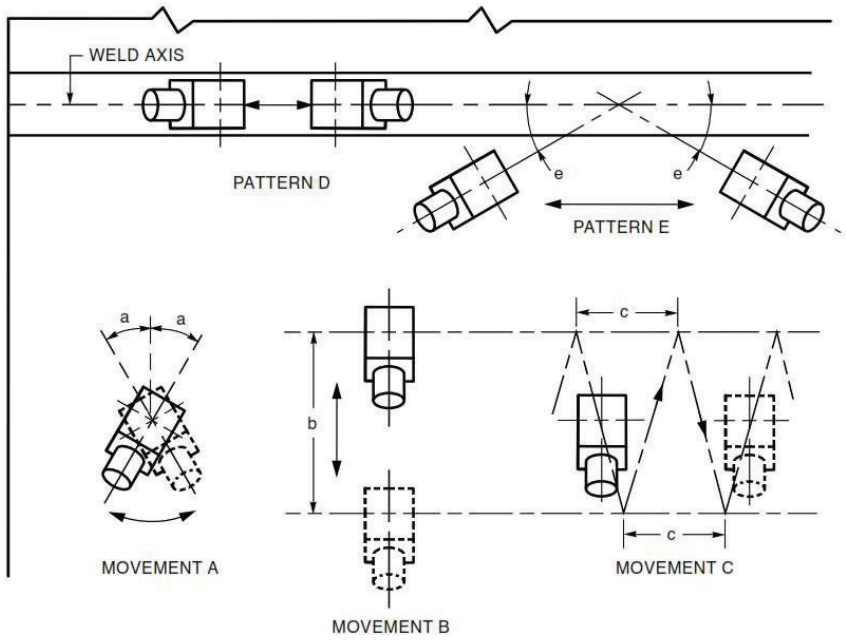


Illustration of Required Scan Patterns

7.2.2.2. Scanning may be performed at primary reference level sensitivity plus the 6 dB correction factor provided soft gain or color palette alterations are made during evaluation to aid in detection. If scanning is performed at primary reference level plus correction factor, soft gain shall be increased by 6 dB, or the color palette adjusted to end at 50% screen height during the evaluation of the weld data. If color palette adjustment or soft gain increase is not used, 6 dB of additional gain over the correction factor and primary reference level shall be applied during scanning.

7.2.2.3. All Flaws > DRL shall be evaluated and accepted or rejected in accordance with Table 1 below.

TABLE 1
ACCEPTANCE CRITERIA

MAXIMUM DISCONTINUITY AMPLITUDE LEVEL		MAXIMUM DISCONTINUITY LENGTH
CLASS A	(> ARL)	None Allowed
CLASS B	(> SSL, ≤ ARL)	0.50 inch
CLASS C	(> DRL, ≤ SSL)	Middle Half of Weld: 2 inches Top or Bottom Quarter of Weld: 0.75 inch
CLASS D	(≤ DRL)	Disregard

8. FLAW CHARACTERIZATION

8.1. All rejectable flaws shall be characterized.

8.1.1. The type, position and orientation of a flaw is an important factor when performing height sizing and considering repair procedures.

8.1.2. All of the different methods and displays contribute to flaw characterization. Flaw characterization should begin with analysis of the a-scan presentation.

8.1.3. Flaws are generally characterized as planar or volumetric, surface connected or embedded. These categories further subdivided into flaw types. Common weld related flaws include inclusions, porosity, cracks, lack of fusion, and lack of penetration. At a minimum, technicians are required to characterize flaws as planar or volumetric.

8.1.4. Planar flaws are generally considered two-dimensional (length and height) and include cracks, lack of fusion, and lack of penetration. Planar flaws generally reflect more ultrasonic energy, have a short rise/fall echo dynamic (narrow time-base), and have stronger diffracted signals.

8.1.5. Volumetric flaws are generally considered three-dimensional (length, width, and height) and include inclusions and porosity. Volumetric flaws generally reflect less ultrasonic energy and have a longer rise/fall echo dynamic (broad time-base).

8.1.6. Flaw positioning, orientation and part geometry must be considered when performing flaw characterization.

8.2. At a minimum, signal amplitude, a-scan echo dynamic, time-base duration and flaw location shall be used to characterize flaws.

STEP 1

a. Optimize target flaw signal by search unit positing and selecting refracted angles that are as close to perpendicular incidence as practical.

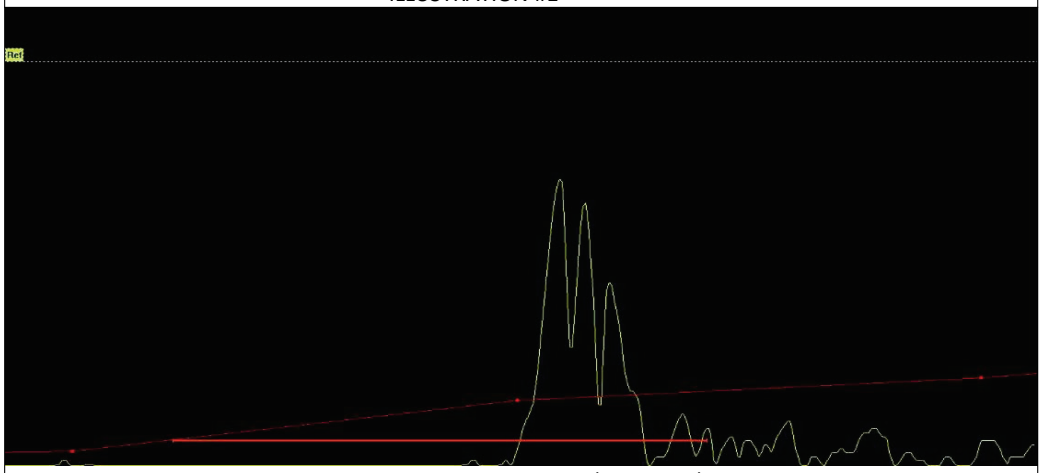
STEP 2

a. Use a cross sectional plot of the weld geometry or a sector scan weld overlay to identify part geometry, flaw position and orientation.

b. Move the search unit towards and away from the target flaw. Observe a-scan amplitude, echo dynamic and time-base duration.

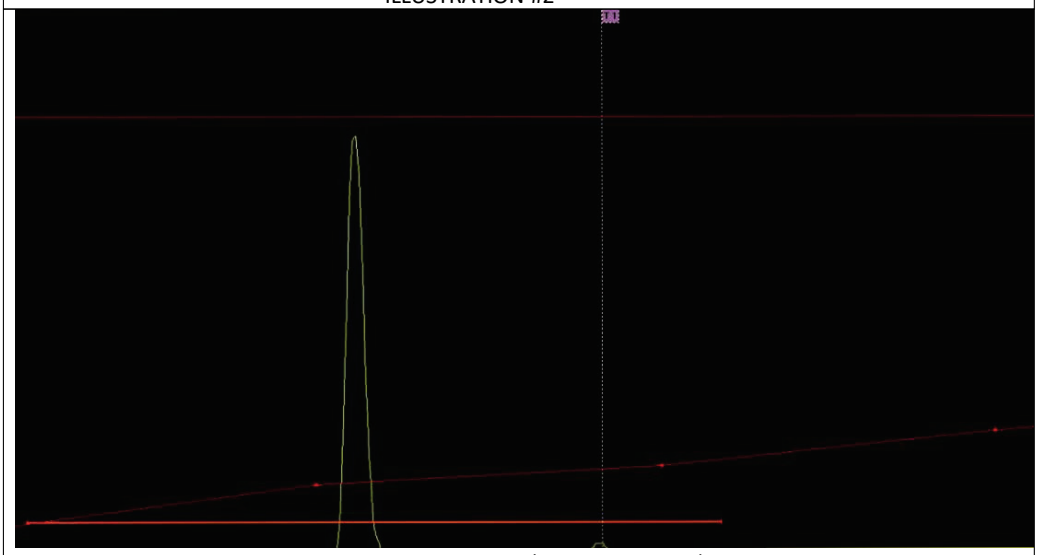
c. With the target flaw signal optimized, rotate search unit. Observe a-scan signal amplitude and fall time.

ILLUSTRATION #1



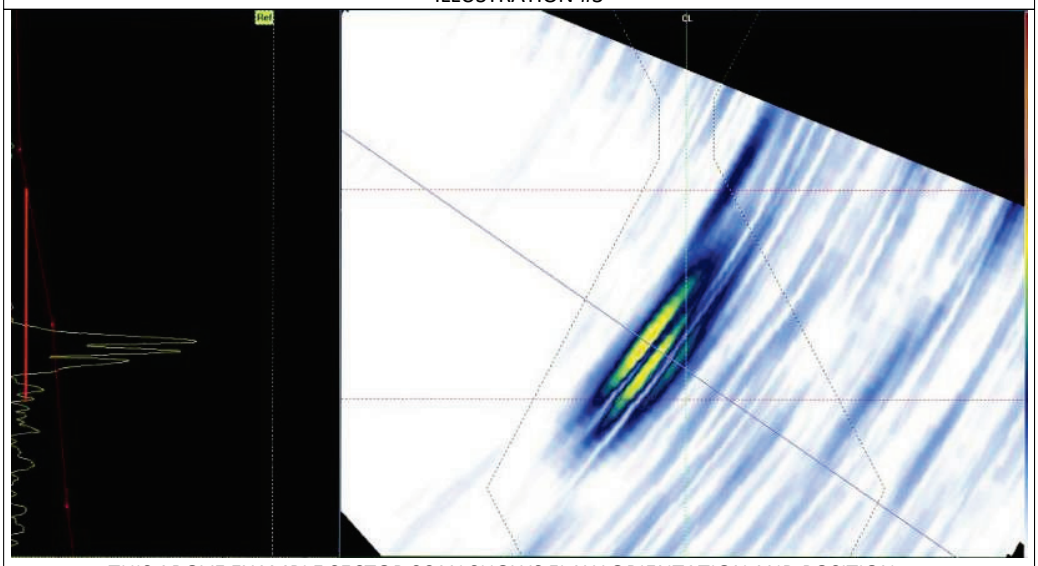
EXAMPLE OF A VOLUMETRIC FLAW (POROSITY)
THE A-SCAN PRESENTATION SHOWS MULTIPLE REFLECTORS, A BROAD TIME-BASE WITH TRAILING SIGNALS

ILLUSTRATION #2



EXAMPLE OF A PLANAR FLAW (LACK OF FUSION)
THE A-SCAN PRESENTATION SHOWS A HIGHER AMPLITUDE SINGLE REFLECTOR, A NARROW TIME-BASE AND NO TRAILING SIGNALS

ILLUSTRATION #3



THIS ABOVE EXAMPLE SECTOR SCAN SHOWS FLAW ORIENTATION AND POSITION

9. LENGTH SIZING

9.1. The 6 dB drop method shall be used for length sizing. The scan axis cursors shall be placed at the color palette -6 dB start and stop positions.

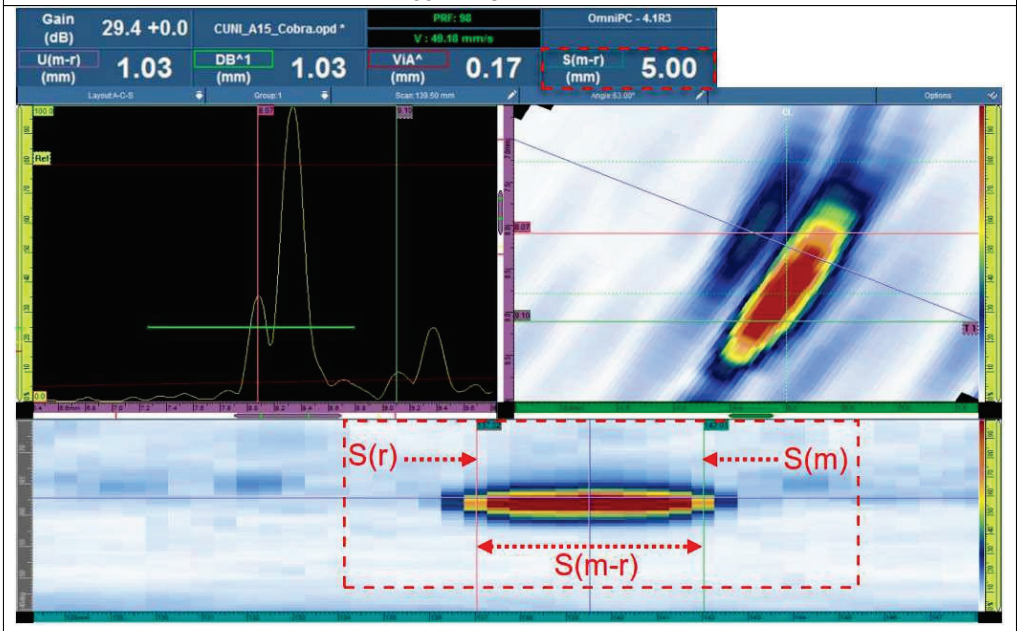
9.2. Please see Illustration #4 of c-scan length sizing.

9.3. The red scan axis reference cursor (r) is positioned at the -6 dB start position (137 mm).

9.4. The green measure cursor (m) is positioned at the -6 dB stop position (142 mm)

9.5. The difference between the scan axis cursors shall be recorded on the ultrasonic test report as flaw length. Please see Illustration #4 below for an example of 6 dB drop length sizing. In this example 142 – 137 = 5 mm flaw length

ILLUSTRATION #4



EXAMPLE OF LENGTH SIZING USING THE 6 DECIBEL DROP METHOD

10. HEIGHT SIZING PROCEDURE (dB Drop)

10.1. The accuracy of flaw height sizing depends on many factors. At a minimum, flaw orientation, flaw type and beam divergence must be considered. Beam profile should be plotted and overlayed to avoid sizing errors.

10.2. Precise calibration is essential for height sizing. Angle, index, delay, and velocity errors all affect height sizing accuracy.

10.3. The 6 dB, 3 dB or rapid dB drop method may be used. The method used shall be recorded on the ultrasonic test report.

STEP 1

a. Recall flaw type recorded on the ultrasonic test report from Section 8

b. Consider flaw type, position and orientation to optimize technique

STEP 2

a. Maximize target flaw signal amplitude by search unit positing and adjusting refracted angles to as close to perpendicular incidence as practical.

b. Record the depth from the maximized signal on the ultrasonic test report

STEP 3

a. Place the ultrasonic scan axis reference cursor on the upper part of the flaw at the desired – dB color palette (e.g. -6 dB, -3 dB etc.).

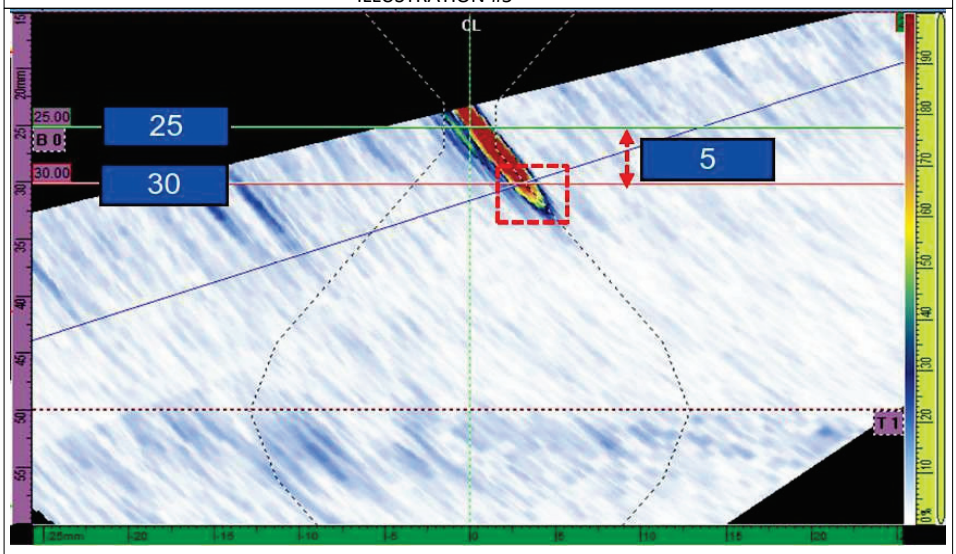
b. Place the ultrasonic scan axis measure cursor on the lower part of the flaw at the desired – dB color palette (e.g. -6 dB, -3 dB etc.).

c. The difference between the ultrasonic scan axis reference cursor and ultrasonic scan axis measure cursor is the estimated flaw height

d. Record this value as flaw height on the ultrasonic test report

10.4. Below is an example of an embedded flaw. This example is a sector scan using the 6 dB drop height sizing method. In this example, the reference cursor is moved to the upper edge of the flaw and positioned at the -6 dB location. The measure cursor is placed at B0. The difference between the reference cursor and the measure cursor is the estimated height. In this example 30 – 25 = 5 mm flaw height.

ILLUSTRATION #5



EXAMPLE OF HEIGHT SIZING USING THE 6 DECIBEL DROP METHOD

11. HEIGHT SIZING PROCEDURE (DIFFRACTION METHOD)

11.1. If diffracted signals are observed, the dB drop methods should not be used. Instead, cursors should be placed at center of energy locations.

11.2. For surface connected flaws, the measure cursor is placed at B0 (face-b) or T1 (face-a) as applicable and the reference cursor placed at the diffracted signal center of energy.

11.3. For embedded flaws, both cursors are placed at signal center of energy locations.

STEP 1

a. Optimize target flaw signal by search unit positing and selecting refracted angles that are as close to perpendicular incidence as practical.

STEP 2

e. Place the ultrasonic scan axis reference cursor on the uppermost diffracted signal center of energy. Record this depth as Z_1 on the ultrasonic test report.

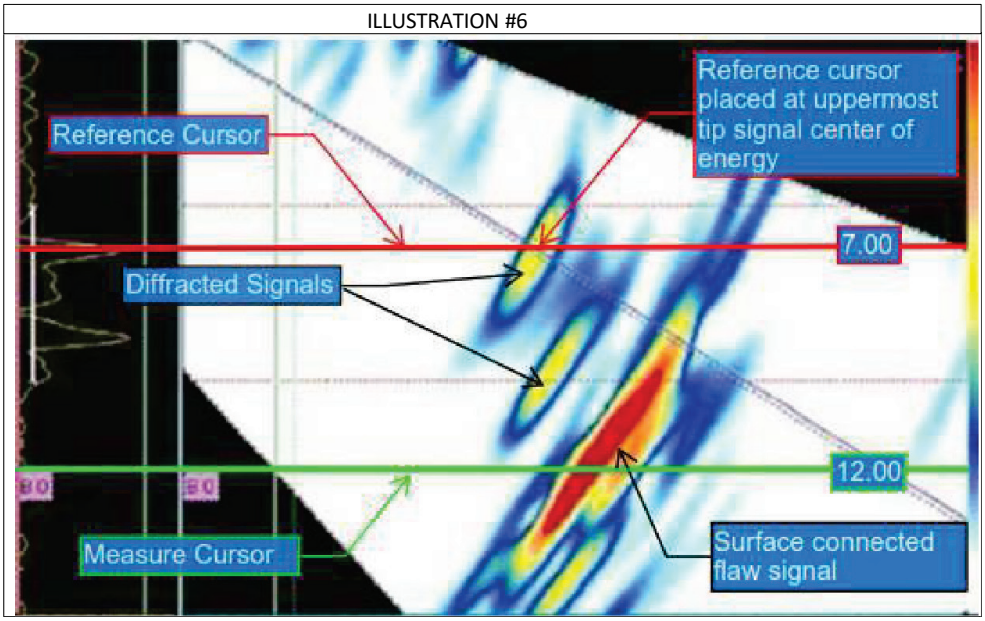
f. For surface connected flaws, place the ultrasonic scan axis measure cursor at the surface where the flaw is connected. Record this depth as Z_2 on the ultrasonic test report.

g. For embedded flaws, place the ultrasonic scan axis reference cursors at the main signal center of energy. Record this depth as Z_2 on the ultrasonic test report.

h. The difference between the reference cursor and measure cursor is the estimated flaw height. Record this value as flaw height on the ultrasonic test report.

11.4. Below is an example of a flaw connected to B0 (face-b) with two diffracted crack tip signals. This example is a sector scan using diffraction sizing techniques. In this example, the measure cursor is moved to material thickness (12 mm). The reference cursor is moved to the center of energy uppermost tip diffracted signal (7 mm). The difference between the reference cursor and measure cursors is the estimated flaw height. In this example 17 – 12 = 5 mm flaw height.

ILLUSTRATION #6



12. FULL MATRIX CAPTURE (FMC) & TOTAL FOCUS MATRIX (TFM)

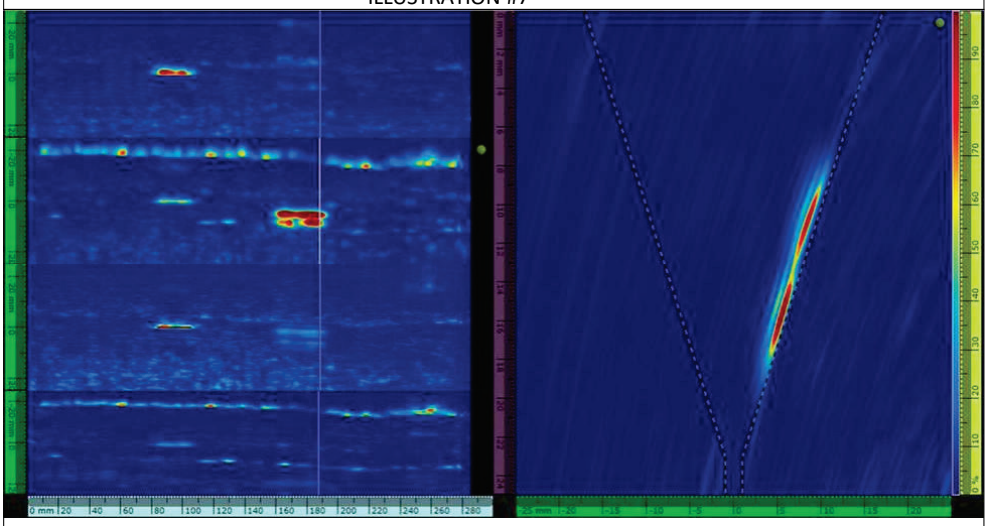
12.1. Ultrasonic modelling software such as Acoustic Influence Map (AIM) or CIVA should be used to select the best wave mode/sets for the target flaws in question.

12.2. It is recommended that the self-tandem mode using TTTTT or TTT be used to start.

12.3. TT, TLT & TTTT wave sets may be used to compliment TTTTT/TTT and are often beneficial for detecting surface connected flaws and diffracted signals.

12.4. Below is an illustrating showing improved imaging of lack of fusion on the weld bevel using TFM/FMC.

ILLUSTRATION #7



13. REPORTING

13.1. A separate ultrasonic test report shall be used for each individual test sample

13.2. All rejected flaws shall be reported on the ultrasonic test report and include the following information:

a. Technicians Name
b. Date
c. Specimen Identification
d. Flaw Number
e. Testing Method (PAUT/TFM/FMC)
f. Scanning Face & Side
g. Primary Reference
h. Flaw Depth
i. Flaw Type
j. Flaw Class
k. Sizing Method
l. Flaw Heightp
m. Flaw Length
n. Distance From Y
o. Distance From X

TECHNICIANS NAME: 13.2 a DATE: 13.2 b SPECIMEN ID: 13.2 c

FLAW NUMBER	TESTING METHOD	SCANNING FACE & SIDE	PRIMARY REFERENCE	TOP OF FLAW (Z_1)	BOTTOM OF FLAW (Z_2)	FLAW HEIGHT ($Z_1 - Z_2$)	FLAW TYPE	FLAW CLASS	FLAW LENGTH	DISTANCE FROM Y	DISTANCE FROM X	REMARKS
13.2 d	13.2 e	13.2 f	6.8	10 11	10 11	10 11	8	7.2.2.3 Table 1	9	7.1	7.2	

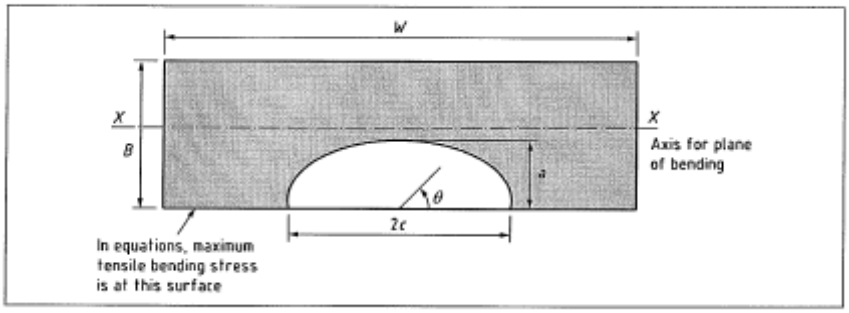
Appendix F: Complete Calculations for Fitness-for-Service (FFS) Examples

This appendix contains the complete calculations for fitness-for-service (FFS) examples.

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

The following spreadsheet uses BS7910:2015 to perform an Option 1 fitness for service assessment of a surface flaw in a plate (Ref. Fig. M.3 from BS7910).

Figure M.3 Surface flaw



User Inputs

$height := 0.2 \cdot in$	Reported crack height
$length := 0.6666 \cdot in$	Reported crack length
$W := 12\ in$	Plate width
$B := 0.75 \cdot in$	Plate thickness
$P_m := 15 \cdot ksi$	Applied membrane stress
$P_b := 0 \cdot ksi$	Applied bending stress
$Q_m := 44 \cdot ksi$	Residual membrane stress
$Q_b := 0 \cdot ksi$	Residual bending stress
$\sigma_y := 44 \cdot ksi$	Material yield strength
$\sigma_u := 78 \cdot ksi$	Material ultimate strength
$K_{mat} := 60 \cdot ksi \cdot \sqrt{in}$	Material toughness (for flange)
$E := 29000 \cdot ksi$	Material elastic modulus
$\theta := 90 \cdot deg$	Parametric angle
$FOS_{height} := 1$	Partial Safety Factor - Flaw height
$FOS_{length} := 1$	Partial Safety Factor - Flaw length

$a := height \cdot FOS_{height} = 0.2\ in$ Length of crack front assumed for analysis

$c := \frac{length}{2} \cdot FOS_{length} = 0.3333\ in$ Half of flaw width assumed for analysis

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

LOAD RATIO:

Determine the reference stress, σref, in accordance with Annex P6

$$\alpha(a,c) := \begin{cases} \text{if } W \geq 2\cdot(c+B) \\ \quad \dfrac{\left(\frac{a}{B}\right)}{1+\left(\frac{B}{c}\right)} \\ \text{else} \\ \quad \left(\left(\frac{a}{B}\right)\cdot\left(\frac{2\cdot c}{W}\right)\right) \end{cases}$$

Assume normal bending restraint = 1

$$\sigma_{ref}(a,c) := \frac{P_b + \left(P_b{}^2 + 9\cdot P_m{}^2\cdot\left(1-\alpha(a,c)\right)^2\right)^{0.5}}{3\cdot\left(1-\alpha(a,c)\right)^2} \qquad \text{(Eq. P.9)}$$

$\sigma_{ref}(a,c) = 16.341$ ***ksi***

Calculate the load ratio, Lr, in accordance with Section 7.3.7

$$L_r(a,c) := \frac{\sigma_{ref}(a,c)}{\sigma_y} \qquad \text{(Eq. 40)}$$

$L_r(a,c) = 0.371$

NOTE: NASGRO USES A SLIGHTLY DIFFERENT DEFINITION OF LOAD RATIO

Determine the plastic load limit, NL, in accordance with NASGRO Appendix X.5.7

$$\alpha_{FAD} := \frac{a}{B} = 0.267 \qquad \beta_{FAD} := \frac{2\cdot c}{W} = 0.056 \qquad \psi_{FAD} := \frac{c}{B} = 0.444 \qquad \lambda_{FAD} := \frac{P_b}{6\cdot P_m} = 0$$

$$d_1 := \left(1-\alpha_{FAD}\cdot\beta_{FAD}\right)^2 + 2\cdot\alpha_{FAD}{}^2\cdot\beta_{FAD}\cdot\left(1-\beta_{FAD}\right) = 0.978$$

$$L_{rN} := \frac{d_1}{2\cdot\lambda_{FAD} + \alpha_{FAD}\cdot\beta_{FAD} + \sqrt{\left(2\cdot\lambda_{FAD} + \left(\alpha_{FAD}\cdot\beta_{FAD}\right)\right)^2 + d_1}} = 0.974$$

$$L_{r\_NG} := \frac{P_m}{L_{rN}\cdot\sigma_y} = 0.3499$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

FRACTURE RATIO:

Determine stress intensity factor, KI, in accordance with Annex M.4

$M := 1$

$$Check := \left\| \begin{array}{l} \text{if } \frac{2 \cdot c}{W} < 0.8 \quad = \text{“Procced”} \\ \quad \left\| \text{“Procced”} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$f_w(a, c) := \left(sec\left(\frac{\pi \cdot c}{W} \cdot \left(\frac{a}{B} \right)^{0.5} \right) \right)^{0.5}$$

Stress intensity magnification for membrane loading (M.4.1.2)

$$Check_1 := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \leq 1.0 \quad = \text{“Proceed”} \\ \quad \left\| \text{“Proceed”} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$Check_2 := \left\| \begin{array}{l} \text{if } 0 < \theta \leq \pi \quad = \text{“Procced”} \\ \quad \left\| \text{“Procced”} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$Check_3 := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \leq 0.1 \quad = \text{“Proceed”} \\ \quad \left\| \begin{array}{l} \text{if } \frac{a}{B} < 1.25 \cdot \left(\frac{a}{c} + 0.6 \right) \\ \quad \left\| \text{“Proceed”} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right. \\ \text{also if } \frac{a}{B} < 1.0 \\ \quad \left\| \text{“Proceed”} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

$$M_1(a,c) := \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad 1.13 - 0.09 \cdot \frac{a}{c} \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left(\frac{c}{a}\right)^{0.5} \cdot \left(1 + 0.04 \cdot \frac{c}{a}\right) \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{array}$$

$$M_2(a,c) := \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left(\frac{0.89}{0.2 + \left(\frac{a}{c}\right)}\right) - 0.54 \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad 0.2 \cdot \left(\frac{c}{a}\right)^4 \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{array}$$

$$M_3(a,c) := \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad 0.5 - \frac{1}{0.65 + \frac{a}{c}} + 14 \cdot \left(1 - \frac{a}{c}\right)^{24} \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad -0.11 \cdot \left(\frac{c}{a}\right)^4 \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{array}$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

$$g(a,c) := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.5 \\ \quad 1 + \left(0.1 + 0.35 \cdot \left(\dfrac{a}{B}\right)^2\right) \cdot (1 - \sin(\theta))^2 \\ \text{also if } 0.5 < \dfrac{a}{2 \cdot c} \le 1.0 \\ \quad 1 + \left(0.1 + 0.35 \cdot \left(\dfrac{c}{a}\right) \cdot \left(\dfrac{a}{B}\right)^2\right) \cdot (1 - \sin(\theta))^2 \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{cases}$$

$$f_\theta(a,c) := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.5 \\ \quad \left(\left(\dfrac{a}{c}\right)^2 \cdot \cos(\theta)^2 + \sin(\theta)^2\right)^{0.25} \\ \text{also if } 0.5 < \dfrac{a}{2 \cdot c} \le 1.0 \\ \quad \left(\left(\dfrac{c}{a}\right)^2 \cdot \sin(\theta)^2 + \cos(\theta)^2\right)^{0.25} \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{cases}$$

$$\Phi(a,c) := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.5 \\ \quad \left(1 + 1.464 \cdot \left(\dfrac{a}{c}\right)^{1.65}\right)^{0.5} \\ \text{also if } 0.5 < \dfrac{a}{2 \cdot c} \le 1.0 \\ \quad \left(1 + 1.464 \cdot \left(\dfrac{c}{a}\right)^{1.65}\right)^{0.5} \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{cases}$$

$$M_m(a,c) := \left(M_1(a,c) + M_2(a,c) \cdot \left(\frac{a}{B}\right)^2 + M_3(a,c) \cdot \left(\frac{a}{B}\right)^4\right) \cdot \frac{g(a,c) \cdot f_\theta(a,c)}{\Phi(a,c)} \qquad \text{(Eq. M.9)}$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

Stress intensity magnification for bending (M.4.1.3)

$$q(a,c) := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left\| 0.2 + \frac{a}{c} + 0.6 \cdot \frac{a}{B} \right. \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left\| 0.2 + \frac{c}{a} + 0.6 \cdot \frac{a}{B} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$H_I(a,c) := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left\| 1 - 0.34 \cdot \left(\frac{a}{B}\right) - 0.11 \cdot \left(\frac{a}{c}\right) \cdot \left(\frac{a}{B}\right) \right. \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left\| 1 - \left(0.04 + 0.41 \cdot \left(\frac{c}{a}\right)\right) \cdot \left(\frac{a}{B}\right) + \left(0.55 - 1.93 \cdot \left(\frac{c}{a}\right)^{0.75} + 1.38 \cdot \left(\frac{c}{a}\right)^{1.5}\right) \cdot \left(\frac{a}{B}\right)^2 \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$G_I(a,c) := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left\| -1.22 + -0.12 \cdot \left(\frac{a}{c}\right) \right. \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left\| -2.11 + 0.77 \cdot \left(\frac{c}{a}\right) \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

$$G_2(a,c) := \left\| \begin{array}{l} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left\| 0.55 - 1.05 \cdot \left(\frac{a}{c}\right)^{0.75} + 0.47 \cdot \left(\frac{a}{c}\right)^{1.5} \right. \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left\| 0.55 - 0.72 \cdot \left(\frac{c}{a}\right)^{0.75} + 0.14 \cdot \left(\frac{c}{a}\right)^{1.5} \right. \\ \text{else} \\ \quad \left\| \text{“Equations not applicable”} \right. \end{array} \right.$$

$$H_2(a,c) := 1 + G_1(a,c) \cdot \left(\frac{a}{B}\right) + G_2(a,c) \cdot \left(\frac{a}{B}\right)^2$$

$$H(a,c) := H_1(a,c) + \left(H_2(a,c) - H_1(a,c)\right) \cdot \sin(\theta)^{q(a,c)}$$

$$M_b(a,c) := H(a,c) \cdot M_m(a,c) \qquad \text{(Eq. M.11)}$$

Note: These equations do NOT account for in-plane bending.

$M_{km} := 1$ Stress intensity multiplication factor, = 1 when flaw is not in region of local stress concentration (Section M.11)

$M_{kb} := 1$ Stress intensity multiplication factor, = 1 when flaw is not in region of local stress concentration. (Section M.11)

$k_{tm} := 1$ Stress concentration factor for membrane stress (Section 6.4.4)

$k_{tb} := 1$ Stress concentration factor for bending stress (Section 6.4.4)

$k_m := 1$ Stress concentration factor due to misalignment, = 1 for no misalignment (Annex D)

Stress intensity contribution from primary stress (Eq. M.4)

$$Y\sigma_p(a,c) := M \cdot f_w(a,c) \cdot \left(k_{tm} \cdot M_{km} \cdot M_m(a,c) \cdot P_m + k_{tb} \cdot M_{kb} \cdot M_b(a,c) \cdot \left(P_b + \left\langle k_m - 1 \right\rangle \cdot P_m\right)\right)$$

Stress intensity contribution from secondary stress (Eq. M.5)

$$Y\sigma_s(a,c) := M_m(a,c) \cdot Q_m + M_b(a,c) \cdot Q_b$$

Combined stress intensity (Eq. M.3)

$$Y\sigma(a,c) := Y\sigma_p(a,c) + Y\sigma_s(a,c)$$

Calculate maximum stress intensity factor (Eq. M.1)

$$K_I(a,c) := Y\sigma(a,c) \cdot \sqrt{\pi \cdot a}$$

$$K_I(a,c) = 40.851 \ \textit{ksi} \cdot \sqrt{\textit{in}}$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

Determine the plasticity correction factor, ρ, in accordance with Annex R

Stress intensity factor due to secondary stresses only

$$K_{I\_s}(a,c) := Y\sigma_s(a,c)\cdot(\pi\cdot a)^{0.5}$$

$$K_{I\_s}(a,c) = 30.462\ ksi\cdot\sqrt{in}$$

Stress intensity factor due to primary stresses only

$$K_{I\_p}(a,c) := Y\sigma_p(a,c)\cdot(\pi\cdot a)^{0.5}$$

$$K_{I\_p}(a,c) = 10.39\ ksi\cdot\sqrt{in}$$

$$X(a,c) := \frac{K_{I\_s}(a,c)\cdot L_r(a,c)}{K_{I\_p}(a,c)}$$

$$\rho_1(a,c) := \begin{cases} \text{if } X(a,c) < 5.2 \\ \quad \left(0.1\cdot(X(a,c))^{0.714} - 0.007\cdot(X(a,c))^{2} + 0.00003\cdot(X(a,c))^{5}\right) \\ \text{else} \\ \quad 0.25 \end{cases} \qquad \text{(Eq. R.1)}$$

$$\rho(a,c) := \begin{cases} \text{if } L_r(a,c) \le 0.8 \\ \quad \rho_1(a,c) \\ \text{also if } L_r(a,c) \ge 1.05 \\ \quad 0 \\ \text{else} \\ \quad \left(4\cdot\rho_1(a,c)\cdot\left(1.05 - L_r(a,c)\right)\right) \end{cases} \qquad \text{(Eq. R.2)}$$

$$\rho(a,c) = 0.098$$

Calculate the Fracture Ratio, Kr, in accordance with Section 7.3.6

$$K_r(a,c) := \frac{K_I(a,c)}{K_{mat}} + \rho(a,c) \qquad \text{(Eq. 39)}$$

$$K_r(a,c) = 0.779$$

NOTE: NASGRO ESTIMATES THE PLASTICITY CORRECTION FACTOR BASED ON INTERPOLATION FROM TABLES

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

Determine the plasticity correction factor in accordance with NASGRO Appendix X.3.2.1

$\beta_{NASGRO} := 3$ (in plane strain)

$K_{I\_s\_NASGRO} := K_{I\_s}(a, c) = 30.462\ ksi \cdot \sqrt{in}$

$K_{I\_p\_NASGRO} := K_{I\_p}(a, c) = 10.39\ ksi \cdot \sqrt{in}$

$$a_{eff} := a + \frac{1}{2 \cdot \pi \cdot \beta_{NASGRO}} \cdot \left(\frac{K_{I\_s\_NASGRO}}{\sigma_y} \right)^2 = 0.225\ in$$

$$K_{I\_s\_plastic} := \sqrt{\frac{a_{eff}}{a}} \cdot K_{I\_s\_NASGRO} = 32.34\ ksi \cdot \sqrt{in}$$

$$K := \frac{K_{I\_s\_plastic} \cdot L_{r\_NG}}{K_{I\_p\_NASGRO}} = 1.089$$

Interpolate ψ and ϕ from Table X.2 and Table X.3

$$\psi := \begin{bmatrix} 0.057 & 0.085 \\ 0.065 & 0.094 \end{bmatrix} \qquad X := \begin{bmatrix} 0.3 & 1.0 \\ 0.4 & 1.5 \end{bmatrix} \qquad \phi := \begin{bmatrix} 0.715 & 0.752 \\ 0.651 & 0.696 \end{bmatrix}$$

$$\psi_1 := \psi_{0,0} - \left(\frac{\psi_{0,0} - \psi_{1,0}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.061 \qquad \phi_1 := \phi_{0,0} - \left(\frac{\phi_{0,0} - \phi_{1,0}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.683$$

$$\psi_2 := \psi_{0,1} - \left(\frac{\psi_{0,1} - \psi_{1,1}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.089 \qquad \phi_2 := \phi_{0,1} - \left(\frac{\phi_{0,1} - \phi_{1,1}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.724$$

$$\Psi_{NG} := \psi_2 - \left(\frac{\psi_2 - \psi_1}{X_{1,1} - X_{0,1}} \cdot \left(X_{1,1} - K \right) \right) = 0.066 \qquad \phi_{NG} := \phi_2 - \left(\frac{\phi_2 - \phi_1}{X_{1,1} - X_{0,1}} \cdot \left(X_{1,1} - K \right) \right) = 0.69$$

$$\rho_{NG} := \Psi_{NG} - \phi_{NG} \cdot \left(\frac{K_{I\_s\_NASGRO}}{K_{I\_s\_plastic}} - 1 \right) = 0.106$$

$$K_{r\_NASGRO} := \frac{K_{I\_s\_NASGRO}}{K_{mat}} + \frac{K_{I\_p\_NASGRO}}{K_{mat}} + \rho_{NG} = 0.787$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack

FAILURE ASSESSMENT DIAGRAM (FAD):

Plot the Failure Assessment Diagram in accordance with Section 7.3 (continuous yielding)

$$L_{rmax} := \frac{\sigma_y + \sigma_u}{2 \cdot \sigma_y} = 1.386$$ (Eq. 25)

$$\mu := min\left(0.001 \cdot \frac{E}{\sigma_y}, 0.6\right) = 0.6$$

$$N := 0.3 \cdot \left(1 - \frac{\sigma_y}{\sigma_u}\right) = 0.131$$

$$K_{r1}(L_{r1}) := \begin{cases} \text{if } L_{r1} \le 1 \\ \quad \left(\left(1 + \frac{1}{2} \cdot L_{r1}^{\,2}\right)^{-0.5} \cdot \left(0.3 + 0.7 \cdot \exp\left(-\mu \cdot L_{r1}^{\,6}\right)\right)\right) \\ \text{if } 1 < L_{r1} \wedge L_{r1} < L_{rmax} \\ \quad \left(\left(\frac{3}{2}\right)^{-0.5} \cdot \left(0.3 + 0.7 \cdot \exp(-\mu)\right)\right) \cdot L_{r1}^{\frac{N-1}{2 \cdot N}} \\ \text{if } L_{r1} \ge L_{rmax} \\ \quad 0 \end{cases}$$ (Eqs. 26-28)

$FOS_{height} := 3$ Partial Safety Factor - Flaw height

$FOS_{length} := 2$ Partial Safety Factor - Flaw length

Crack 2: Reported height = 0.2 in., Reported length = 0.23 in

$a_2 := 0.2 \cdot in$ Length of crack front assumed for analysis

$c_2 := 0.23\ in$ Half of flaw width assumed for analysis

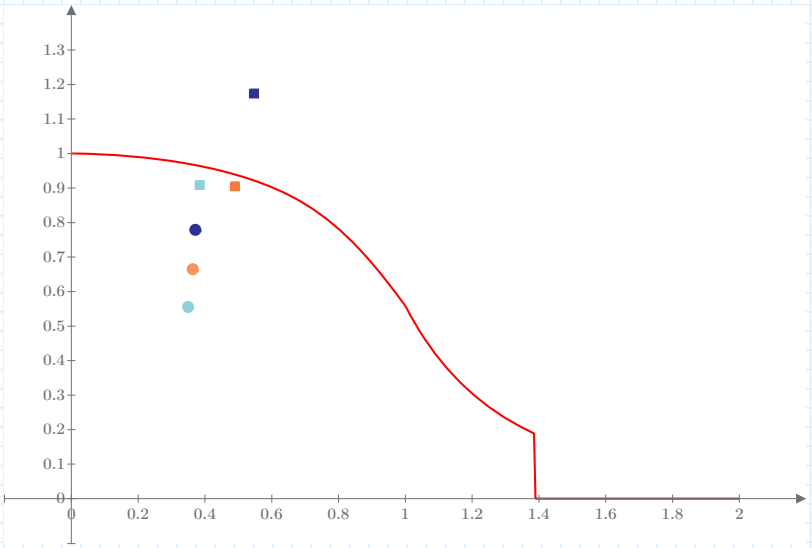
Crack 3: Reported height = 0.06 in., Reported length = 0.66 in

$a_3 := 0.06 \cdot in$ Length of crack front assumed for analysis

$c_3 := 0.3333 \cdot in$ Half of flaw width assumed for analysis

$$a_{all} := \begin{bmatrix} a \\ a \cdot FOS_{height} \\ a_2 \\ a_2 \cdot FOS_{height} \\ a_3 \\ a_3 \cdot FOS_{height} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.6 \\ 0.2 \\ 0.6 \\ 0.06 \\ 0.18 \end{bmatrix} in \qquad c_{all} := \begin{bmatrix} c \\ c \cdot FOS_{length} \\ c_2 \\ c_2 \cdot FOS_{length} \\ c_3 \\ c_3 \cdot FOS_{length} \end{bmatrix} = \begin{bmatrix} 0.333 \\ 0.667 \\ 0.23 \\ 0.46 \\ 0.333 \\ 0.667 \end{bmatrix} in$$

Example 1 - Fracture Evaluation of Flat Plate with Surface Crack



$K_{r1}(L_{r1})$

$K_r\left(a_{all_0}, c_{all_0}\right)$

$K_r\left(a_{all_1}, c_{all_1}\right)$

$K_r\left(a_{all_2}, c_{all_2}\right)$

$K_r\left(a_{all_3}, c_{all_3}\right)$

$K_r\left(a_{all_4}, c_{all_4}\right)$

$K_r\left(a_{all_5}, c_{all_5}\right)$

L_{r1}

$L_r\left(a_{all_0}, c_{all_0}\right)$

$L_r\left(a_{all_1}, c_{all_1}\right)$

$L_r\left(a_{all_2}, c_{all_2}\right)$

$L_r\left(a_{all_3}, c_{all_3}\right)$

$L_r\left(a_{all_4}, c_{all_4}\right)$

$L_r\left(a_{all_5}, c_{all_5}\right)$

$K_r\left(a_{all_0}, c_{all_0}\right) = 0.779$

$K_r\left(a_{all_1}, c_{all_1}\right) = 1.174$

$K_r\left(a_{all_2}, c_{all_2}\right) = 0.665$

$K_r\left(a_{all_3}, c_{all_3}\right) = 0.905$

$K_r\left(a_{all_4}, c_{all_4}\right) = 0.555$

$K_r\left(a_{all_5}, c_{all_5}\right) = 0.909$

$L_r\left(a_{all_0}, c_{all_0}\right) = 0.371$

$L_r\left(a_{all_1}, c_{all_1}\right) = 0.547$

$L_r\left(a_{all_2}, c_{all_2}\right) = 0.364$

$L_r\left(a_{all_3}, c_{all_3}\right) = 0.49$

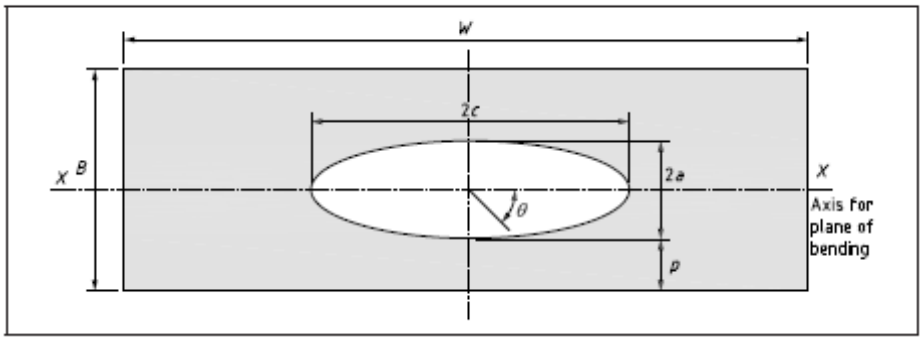
$L_r\left(a_{all_4}, c_{all_4}\right) = 0.35$

$L_r\left(a_{all_5}, c_{all_5}\right) = 0.384$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

The following spreadsheet uses BS7910:2015 to perform an Option 1 fitness for service assessment of an embedded flaw in a plate (Ref. Fig. M.8 from BS7910).

Figure M.8 **Embedded flaw**



User Inputs

$height := 0.2 \cdot in$	Reported crack height
$length := 0.6666 \cdot in$	Reported crack length
$W := 12\ in$	Plate width
$B := 0.75 \cdot in$	Plate thickness
$P_m := 15 \cdot ksi$	Applied membrane stress
$P_b := 5 \cdot ksi$	Applied bending stress
$Q_m := 36 \cdot ksi$	Residual membrane stress
$Q_b := 0 \cdot ksi$	Residual bending stress
$\sigma_y := 36 \cdot ksi$	Material yield strength
$\sigma_u := 58 \cdot ksi$	Material ultimate strength
$K_{mat} := 60 \cdot ksi \cdot \sqrt{in}$	Material toughness (for flange)
$E := 29000 \cdot ksi$	Material elastic modulus
$\theta := 90\ deg$	Parametric angle
$FOS_{height} := 1$	Partial Safety Factor - Flaw height
$FOS_{length} := 1$	Partial Safety Factor - Flaw length

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

$a := \frac{height}{2} \cdot FOS_{height} = 0.1\ in$ Half of crack front length assumed for analysis

$c := \frac{length}{2} \cdot FOS_{length} = 0.3333\ in$ Half of flaw width assumed for analysis

$p(a) := \frac{B - 2 \cdot a}{2}$ Clear distance from flaw to edge (assume flaw is centered)

$B_p(a) := 2 \cdot a + 2 \cdot p(a)$ Effective width

LOAD RATIO:

Determine the reference stress, σref, in accordance with Annex P6

$$\alpha(a,c) := \begin{cases} \text{if } W \geq 2 \cdot (c + B) \\ \quad \dfrac{\left(\frac{2\,a}{B}\right)}{1 + \left(\frac{B}{c}\right)} \\ \text{else} \\ \quad \left(\left(\frac{2\,a}{B}\right) \cdot \left(\frac{2 \cdot c}{W}\right)\right) \end{cases}$$

$$\sigma_{ref}(a,c) := \frac{P_b + 3 \cdot P_m \cdot \alpha(a,c)}{3 \cdot \left(\left(1 - \alpha(a,c)\right)^2 + 4 \cdot \left(\frac{p(a) \cdot \alpha(a,c)}{B}\right)\right)} + \frac{\left(\left(P_b + 3 \cdot P_m \cdot \alpha(a,c)\right)^2 + 9 \cdot {P_m}^2 \cdot \left(\left(1 - \alpha(a,c)\right)^2 + 4 \cdot \left(\frac{p(a) \cdot \alpha(a,c)}{B}\right)\right)\right)^{0.5}}{3 \cdot \left(\left(1 - \alpha(a,c)\right)^2 + 4 \cdot \left(\frac{p(a) \cdot \alpha(a,c)}{B}\right)\right)} \quad \text{(Eq. P.11)}$$

$\sigma_{ref}(a,c) = 18.588\ ksi$

Calculate the load ratio, Lr, in accordance with Section 7.3.7

$L_r(a,c) := \frac{\sigma_{ref}(a,c)}{\sigma_y}$ (Eq. 40)

$L_r(a,c) = 0.516$

NOTE: NASGRO USES A SLIGHTLY DIFFERENT DEFINITION OF LOAD RATIO

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

Determine the plastic load limit, NL, in accordance with NASGRO Appendix X.5.8

$B_w := \frac{B}{2} = 0.375$ *in* $\quad \alpha_{FAD} := \frac{a}{B} = 0.133 \quad \beta_{FAD} := \frac{2 \cdot c}{W} = 0.056 \quad \psi_{FAD} := \frac{c}{B} = 0.444$

$\lambda_{FAD} := \frac{P_b}{6 \cdot P_m} = 0.056 \quad k := \frac{1}{2} - \frac{B_w}{B} = 0$

$c_1 := 1 - 8 \cdot \alpha_{FAD} \cdot \beta_{FAD} \cdot k - 4 \cdot (\alpha_{FAD} \cdot \beta_{FAD})^2 = 1$

$c_2 := (1 - \beta_{FAD}) \cdot \left(1 - \frac{4 \cdot \beta_{FAD} \cdot k^2}{1 - \beta_{FAD}} - 4 \cdot \beta_{FAD} \cdot \alpha_{FAD}^2\right) = 0.941$

$\alpha_1 := (k - \lambda_{FAD}) \cdot (1 - \beta_{FAD}) + \sqrt{(k - \lambda_{FAD})^2 \cdot (1 - \beta_{FAD})^2 + \left(\frac{1}{4} - k^2 + 2 \cdot k \cdot \lambda_{FAD}\right)} = 0.45$

$a_2 := \frac{1}{2} - k = 0.5$

$L_{rN} := \frac{c_1}{2 \cdot (\lambda_{FAD} + \alpha_{FAD} \cdot \beta_{FAD}) + \sqrt{4 \cdot (\lambda_{FAD} + \alpha_{FAD} \cdot \beta_{FAD})^2 + c_1}} = 0.882$

$N_L := L_{rN} \cdot W \cdot B \cdot \sigma_y = 285.724$ *kip*

$L_{r\_NG} := \frac{P_m}{L_{rN} \cdot \sigma_y} = 0.4725$ (The same load ratio is calculated using the bending stress)

$L_r(a, c)$

FRACTURE RATIO:

Determine stress intensity factor, KI, in accordance with Annex M.4

$M := 1$

$$Check := \begin{cases} \text{if } \frac{2 \cdot c}{W} < 0.8 \\ \quad \text{"Procced"} \\ \text{else} \\ \quad \text{"Equations not applicable"} \end{cases} = \text{"Procced"}$$

$$f_w(a, c) := \left(sec\left(\frac{\pi \cdot c}{W} \cdot \left(\frac{2\ a}{B}\right)^{0.5}\right)\right)^{0.5}$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

Stress intensity magnification for membrane loading (M.4.3.2)

$$Check_1 := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 1.0 & = \text{“Proceed”} \\ \quad \text{“Proceed”} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

$$Check_2 := \begin{cases} \text{if } -\pi < \theta \le \pi & = \text{“Procced”} \\ \quad \text{“Procced”} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

$$Check_3 := \begin{cases} \text{if } 0 < \dfrac{2 \cdot c}{W} \le 0.5 & = \text{“Proceed”} \\ \quad \text{“Proceed”} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

$$Check_4 := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.1 & = \text{“N/A”} \\ \quad \text{if } \dfrac{a}{B_p} < 0.625 \cdot \left(\dfrac{a}{c} + 0.6 \right) & \\ \quad\quad \text{“Proceed”} & \\ \quad \text{else} & \\ \quad\quad \text{“Equations not applicable”} & \\ \text{else} & \\ \quad \text{“N/A”} & \end{cases}$$

$$M_1(a, c) := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.5 & \\ \quad 1 & \\ \text{also if } 0.5 < \dfrac{a}{2 \cdot c} \le 1.0 & \\ \quad \left(\dfrac{c}{a} \right)^{0.5} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

$$M_2(a,c) := \frac{0.05}{0.11 + \left(\frac{a}{c}\right)^{1.5}}$$

$$M_3(a,c) := \frac{0.29}{0.23 + \left(\frac{a}{c}\right)^{1.5}}$$

$$g(a,c) := 1 - \frac{\left(\left(\frac{2 \cdot a}{B_p(a)}\right)^4 \cdot \left(2.6 - \frac{4 \cdot a}{B_p(a)}\right)\right)^{0.5}}{1 + 4 \cdot \left(\frac{a}{c}\right)} \cdot |\cos(\theta)|$$

$$f_\theta(a,c) := \begin{cases} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left(\left(\frac{a}{c}\right)^2 \cdot \cos(\theta)^2 + \sin(\theta)^2\right)^{0.25} \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left(\left(\frac{c}{a}\right)^2 \cdot \sin(\theta)^2 + \cos(\theta)^2\right)^{0.25} \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{cases}$$

$$\Phi(a,c) := \begin{cases} \text{if } 0 < \frac{a}{2 \cdot c} \le 0.5 \\ \quad \left(1 + 1.464 \cdot \left(\frac{a}{c}\right)^{1.65}\right)^{0.5} \\ \text{also if } 0.5 < \frac{a}{2 \cdot c} \le 1.0 \\ \quad \left(1 + 1.464 \cdot \left(\frac{c}{a}\right)^{1.65}\right)^{0.5} \\ \text{else} \\ \quad \text{“Equations not applicable”} \end{cases} \qquad \text{(Eq. M.10)}$$

$$M_m(a,c) := \left(M_1(a,c) + M_2(a,c) \cdot \left(\frac{2\ a}{B_p(a)}\right)^2 + M_3(a,c) \cdot \left(\frac{2\ a}{B_p(a)}\right)^4\right) \cdot \frac{g(a,c) \cdot f_\theta(a,c)}{\Phi(a,c)} \qquad \text{(Eq. M.14)}$$

$$M_m(a,c) = 0.928$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

Stress intensity magnification for bending (M.4.3.3)

$$Check := \begin{cases} \text{if } 0 < \dfrac{a}{2 \cdot c} \le 0.5 & = \text{“Procced”} \\ \quad \text{“Procced”} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

$$Check := \begin{cases} \text{if } \theta = \dfrac{\pi}{2} & = \text{“Procced”} \\ \quad \text{“Procced”} & \\ \text{else} & \\ \quad \text{“Equations not applicable”} & \end{cases}$$

$$\lambda_1(a) := \begin{cases} \text{if } \dfrac{p(a)}{B} \le 0.1841 & \\ \quad 1.044 & \\ \text{if } \dfrac{p(a)}{B} > 0.1841 & \\ \quad \text{if } \dfrac{a}{B} \le 0.125 & \\ \quad\quad 0.94 & \\ \quad \text{else} & \\ \quad\quad 1.06 & \end{cases} \qquad \lambda_2(a) := \begin{cases} \text{if } \dfrac{p(a)}{B} \le 0.1841 & \\ \quad -2.44 & \\ \text{if } \dfrac{p(a)}{B} > 0.1841 & \\ \quad \text{if } \dfrac{a}{B} \le 0.125 & \\ \quad\quad -1.875 & \\ \quad \text{else} & \\ \quad\quad -2.2 & \end{cases}$$

$$\lambda_3(a) := \begin{cases} \text{if } \dfrac{p(a)}{B} \le 0.1841 & \\ \quad 0 & \\ \text{if } \dfrac{p(a)}{B} > 0.1841 & \\ \quad \text{if } \dfrac{a}{B} \le 0.125 & \\ \quad\quad -0.1146 & \\ \quad \text{else} & \\ \quad\quad -0.6666 & \end{cases} \qquad \lambda_4(a) := \begin{cases} \text{if } \dfrac{p(a)}{B} \le 0.1841 & \\ \quad -3.166 & \\ \text{if } \dfrac{p(a)}{B} > 0.1841 & \\ \quad \text{if } \dfrac{a}{B} \le 0.125 & \\ \quad\quad -1.844 & \\ \quad \text{else} & \\ \quad\quad -0.6666 & \end{cases}$$

$$M_b(a,c) := \frac{\lambda_1(a) + \lambda_2(a) \cdot \left(\dfrac{p(a)}{B}\right) + \lambda_3(a) \cdot \left(\dfrac{a}{B}\right) + \lambda_4(a) \cdot \left(\dfrac{p(a) \cdot a}{B^2}\right)}{\Phi(a,c)} \quad \text{(Eq. M.15)}$$

$$M_b(a,c) = 0.12$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

Note: These equations do NOT account for in-plane bending.

$M_{km} := 1$ Stress intensity multiplication factor, = 1 when flaw is not in region of local stress concentration (Section M.11)

$M_{kb} := 1$ Stress intensity multiplication factor, = 1 when flaw is not in region of local stress concentration. (Section M.11)

$k_{tm} := 1$ Stress concentration factor for membrane stress (Section 6.4.4)

$k_{tb} := 1$ Stress concentration factor for bending stress (Section 6.4.4)

$k_m := 1$ Stress concentration factor due to misalignment, = 1 for no misalignment (Annex D)

Stress intensity contribution from primary stress (Eq. M.4)

$$Y\sigma_p(a,c) := M \cdot f_w(a,c) \cdot \left\langle k_{tm} \cdot M_{km} \cdot M_m(a,c) \cdot P_m + k_{tb} \cdot M_{kb} \cdot M_b(a,c) \cdot \left\langle P_b + \left\langle k_m - 1 \right\rangle \cdot P_m \right\rangle \right\rangle$$

Stress intensity contribution from secondary stress (Eq. M.5)

$$Y\sigma_s(a,c) := M_m(a,c) \cdot Q_m + M_b(a,c) \cdot Q_b$$

Combined stress intensity (Eq. M.3)

$$Y\sigma(a,c) := Y\sigma_p(a,c) + Y\sigma_s(a,c)$$

Calculate maximum stress intensity factor (Eq. M.1)

$$K_I(a,c) := Y\sigma(a,c) \cdot \sqrt{\pi \cdot a}$$

$$K_I(a,c) = 26.862 \ ksi \cdot \sqrt{in}$$

Determine the plasticity correction factor, ρ, in accordance with Annex R

Stress intensity factor due to secondary stresses only

$$K_{I\_s}(a,c) := Y\sigma_s(a,c) \cdot (\pi \cdot a)^{0.5}$$

$$K_{I\_s}(a,c) = 18.721 \ ksi \cdot \sqrt{in}$$

Stress intensity factor due to primary stresses only

$$K_{I\_p}(a,c) := Y\sigma_p(a,c) \cdot (\pi \cdot a)^{0.5}$$

$$K_{I\_p}(a,c) = 8.142 \ ksi \cdot \sqrt{in}$$

$$X(a,c) := \frac{K_{I\_s}(a,c) \cdot L_r(a,c)}{K_{I\_p}(a,c)}$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

$$\rho_1(a,c) := \begin{cases} \text{if } X(a,c) < 5.2 \\ \quad \left(0.1 \cdot (X(a,c))^{0.714} - 0.007 \cdot (X(a,c))^{2} + 0.00003 \cdot (X(a,c))^{5}\right) \\ \text{else} \\ \quad 0.25 \end{cases} \qquad \text{(Eq. R.1)}$$

$$\rho(a,c) := \begin{cases} \text{if } L_r(a,c) \le 0.8 \\ \quad \rho_1(a,c) \\ \text{also if } L_r(a,c) \ge 1.05 \\ \quad 0 \\ \text{else} \\ \quad \langle 4 \cdot \rho_1(a,c) \cdot (1.05 - L_r(a,c)) \rangle \end{cases} \qquad \text{(Eq. R.2)}$$

$\rho_1(a,c) = 0.103$

$\rho(a,c) = 0.103$

Calculate the Fracture Ratio, Kr, in accordance with Section 7.3.6

$$K_r(a,c) := \frac{K_I(a,c)}{K_{mat}} + \rho(a,c) \qquad \text{(Eq. 39)}$$

$K_r(a,c) = 0.551$

NOTE: NASGRO ESTIMATES THE PLASTICITY CORRECTION FACTOR BASED ON INTERPOLATION FROM TABLES

Determine the plasticity correction factor in accordance with NASGRO Appendix X.3.2.1

$\beta_{NASGRO} := 3$ (in plane strain)

$K_{I\_s\_NASGRO} := K_{I\_s}(a,c) = 18.721 \ ksi \cdot \sqrt{in}$

$K_{I\_p\_NASGRO} := K_{I\_p}(a,c) = 8.142 \ ksi \cdot \sqrt{in}$

$$a_{eff} := a + \frac{1}{2 \cdot \pi \cdot \beta_{NASGRO}} \cdot \left(\frac{K_{I\_s\_NASGRO}}{\sigma_y}\right)^2 = 0.114 \ in$$

$$K_{I\_s\_plastic} := \sqrt{\frac{a_{eff}}{a}} \cdot K_{I\_s\_NASGRO} = 20.018 \ ksi \cdot \sqrt{in}$$

$$K := \frac{K_{I\_s\_plastic} \cdot L_{r\_NG}}{K_{I\_p\_NASGRO}} = 1.162 \qquad L_{r\_NG} = 0.472$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

Interpolate ψ and ϕ from Table X.2 and Table X.3

$$\psi := \begin{bmatrix} 0.064 & 0.094 \\ 0.074 & 0.105 \end{bmatrix} \qquad X := \begin{bmatrix} 0.4 & 1.0 \\ 0.5 & 1.5 \end{bmatrix} \qquad \phi := \begin{bmatrix} 0.651 & 0.696 \\ 0.589 & 0.640 \end{bmatrix}$$

$$\psi_1 := \psi_{0,0} - \left(\frac{\psi_{0,0} - \psi_{1,0}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.071 \qquad \phi_1 := \phi_{0,0} - \left(\frac{\phi_{0,0} - \phi_{1,0}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.606$$

$$\psi_2 := \psi_{0,1} - \left(\frac{\psi_{0,1} - \psi_{1,1}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.102 \qquad \phi_2 := \phi_{0,1} - \left(\frac{\phi_{0,1} - \phi_{1,1}}{X_{0,0} - X_{1,0}} \cdot \left(X_{0,0} - L_{r\_NG} \right) \right) = 0.655$$

$$\Psi_{NG} := \psi_2 - \left(\frac{\psi_2 - \psi_1}{X_{1,1} - X_{0,1}} \cdot \left(X_{1,1} - K \right) \right) = 0.081 \qquad \phi_{NG} := \phi_2 - \left(\frac{\phi_2 - \phi_1}{X_{1,1} - X_{0,1}} \cdot \left(X_{1,1} - K \right) \right) = 0.622$$

$$\rho_{NG} := \Psi_{NG} - \phi_{NG} \cdot \left(\frac{K_{I\_s\_NASGRO}}{K_{I\_s\_plastic}} - 1 \right) = 0.122$$

$$K_{r\_NASGRO} := \frac{K_{I\_s\_NASGRO}}{K_{mat}} + \frac{K_{I\_p\_NASGRO}}{K_{mat}} + \rho_{NG} = 0.5692$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

FAILURE ASSESSMENT DIAGRAM (FAD):

Plot the Failure Assessment Diagram in accordance with Section 7.3 (discontinuous yielding)

$$L_{rmax} := \frac{\sigma_y + \sigma_u}{2 \cdot \sigma_y} = 1.306 \qquad \text{(Eq. 25)}$$

$$\Delta\varepsilon := \begin{cases} \text{if } \sigma_y < 1000 \cdot MPa \\ \quad \left(0.0375 \cdot \left(1 - \frac{0.001 \cdot \sigma_y}{MPa}\right)\right) \\ \text{else} \\ \quad \text{“Not Applicable”} \end{cases} = 0.028 \qquad \text{(Eq. 8)}$$

$$\lambda := 1 + \frac{E \cdot \Delta\varepsilon}{\sigma_y} = 23.71 \qquad \text{(Eq. 33)}$$

$$N := 0.3 \cdot \left(1 - \frac{\sigma_y}{\sigma_u}\right) = 0.114$$

$$K_{rl}(L_{rl}) := \begin{cases} \text{if } L_{rl} < 1 \\ \quad \left(\left(1 + \frac{1}{2} \cdot {L_{rl}}^2\right)^{-0.5}\right) \\ \text{if } L_{rl} = 1 \\ \quad \left(\lambda + \frac{1}{2 \cdot \lambda}\right)^{-0.5} \\ \text{if } 1 < L_{rl} \wedge L_{rl} < L_{rmax} \\ \quad \left(\lambda + \frac{1}{2 \cdot \lambda}\right)^{-0.5} \cdot {L_{rl}}^{\frac{N-1}{2 \cdot N}} \\ \text{if } L_{rl} \geq L_{rmax} \\ \quad 0 \end{cases} \qquad \text{(Eqs. 29-32)}$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack

For reference, also plot the FAD for continuous yielding

$$\mu := min\left(0.001 \cdot \frac{E}{\sigma_y}, 0.6\right) = 0.6$$

$$K_{r2}(L_{r2}) := \begin{cases} \text{if } L_{r2} \le 1 \\ \quad \left(\left(1 + \frac{1}{2} \cdot {L_{r2}}^2\right)^{-0.5} \cdot \left(0.3 + 0.7 \cdot \exp\left(-\mu \cdot {L_{r2}}^6\right)\right)\right) \\ \text{if } 1 < L_{r2} \wedge L_{r2} < L_{rmax} \\ \quad \left(\left(\frac{3}{2}\right)^{-0.5} \cdot \left(0.3 + 0.7 \cdot \exp(-\mu)\right)\right) \cdot L_{r2}^{\frac{N-1}{2 \cdot N}} \\ \text{if } L_{r2} \ge L_{rmax} \\ \quad 0 \end{cases}$$

$FOS_{height} := 3$ Partial Safety Factor - Flaw height

$FOS_{length} := 2$ Partial Safety Factor - Flaw length

Crack 2: Reported height = 0.2 in., Reported length = 0.40 in

$a_2 := 0.1 \cdot in$ Half of crack front length assumed for analysis

$c_2 := 0.2\ in$ Half of flaw width assumed for analysis

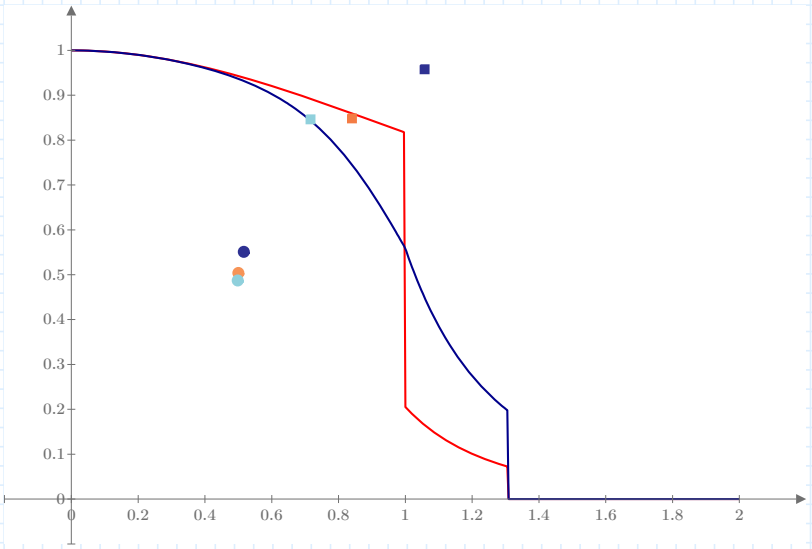
Crack 3: Reported height = 0.14 in., Reported length = 0.66 in

$a_3 := 0.07 \cdot in$ Half of crack front length assumed for analysis

$c_3 := 0.3333 \cdot in$ Half of flaw width assumed for analysis

$$a_{all} := \begin{bmatrix} a \\ a \cdot FOS_{height} \\ a_2 \\ a_2 \cdot FOS_{height} \\ a_3 \\ a_3 \cdot FOS_{height} \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.3 \\ 0.1 \\ 0.3 \\ 0.07 \\ 0.21 \end{bmatrix} in \qquad c_{all} := \begin{bmatrix} c \\ c \cdot FOS_{length} \\ c_2 \\ c_2 \cdot FOS_{length} \\ c_3 \\ c_3 \cdot FOS_{length} \end{bmatrix} = \begin{bmatrix} 0.333 \\ 0.667 \\ 0.2 \\ 0.4 \\ 0.333 \\ 0.667 \end{bmatrix} in$$

Example 2 - Fracture Evaluation of Flat Plate with Embedded Crack



$K_{r1}\left(L_{r1}\right)$

$K_{r2}\left(L_{r2}\right)$

$K_r\left(a_{all_0}, c_{all_0}\right)$

$K_r\left(a_{all_1}, c_{all_1}\right)$

$K_r\left(a_{all_2}, c_{all_2}\right)$

$K_r\left(a_{all_3}, c_{all_3}\right)$

$K_r\left(a_{all_4}, c_{all_4}\right)$

$K_r\left(a_{all_5}, c_{all_5}\right)$

L_{r1}

L_{r2}

$L_r\left(a_{all_0}, c_{all_0}\right)$

$L_r\left(a_{all_1}, c_{all_1}\right)$

$L_r\left(a_{all_2}, c_{all_2}\right)$

$L_r\left(a_{all_3}, c_{all_3}\right)$

$L_r\left(a_{all_4}, c_{all_4}\right)$

$L_r\left(a_{all_5}, c_{all_5}\right)$

$K_r\left(a_{all_0}, c_{all_0}\right) = 0.551$

$K_r\left(a_{all_1}, c_{all_1}\right) = 0.958$

$K_r\left(a_{all_2}, c_{all_2}\right) = 0.504$

$K_r\left(a_{all_3}, c_{all_3}\right) = 0.848$

$K_r\left(a_{all_4}, c_{all_4}\right) = 0.487$

$K_r\left(a_{all_5}, c_{all_5}\right) = 0.846$

$L_r\left(a_{all_0}, c_{all_0}\right) = 0.516$

$L_r\left(a_{all_1}, c_{all_1}\right) = 1.058$

$L_r\left(a_{all_2}, c_{all_2}\right) = 0.5$

$L_r\left(a_{all_3}, c_{all_3}\right) = 0.84$

$L_r\left(a_{all_4}, c_{all_4}\right) = 0.498$

$L_r\left(a_{all_5}, c_{all_5}\right) = 0.716$

FAILURE ASSESSMENT DIAGRAM ANALYSIS
==========
DATE: 26-Mar-23 TIME: 21:40:20

NASGRO(R) Version 10.10 (DLL 64-bit), August 2022
Final Version

PROBLEM TITLE

Ex 2 - Flat Plate with Embedded Crack
U.S. customary units [in, in/cycle, kips, ksi, ksi sqrt(in)]

MODEL: EC05
Shakedown Option: None (or not needed)

Material Data Source: User-entered

GEOMETRY

MODEL: EC05 - Elliptical Embedded Crack (Offset) in Plate – Univariant WF
Thickness, t = 0.7500
Width, W = 12.0000
UNIFORM REMOTE TENSION & REMOTE BENDING SELECTED

USER DEFINED stress gradients:
==========

x	X	S0	x	X	S1
0.000E+00	0.000	0.100E+01	0.000E+00	0.000	0.100E+01
0.750E+00	1.000	0.100E+01	0.750E+00	1.000	-0.100E+01

Optimum point spacing algorithm is NOT enabled. NASGRO uses
the USER-DEFINED point spacing to compute SIFs.

MATERIAL

Material File Name: NASMF.XMLZ
Material File Description: NASA data/NASGRO eqn (single temp)

MATL 1: ASTM SPEC. GRADE STL
A10 series
Material 1, Data ID: B0CB10AB1
Alloy Description: A36 Plt

Alloy Cond/HT:Plt & Sht; ; LA; Room temp

Material Properties:

:Matl: UTS : YS : K1e : K1c : Ak : Bk : Thk : Kc : Keac :
: No.: : : : : : : : :
:----:-------:-------:-------:-------:-------:-------:-------:-------:-------
: 1 : 58.0: 36.0: 60.0: 60.0: 0.75: 0.50: 0.750: 89.8: :

FITNET Option 1 enabled:
Minimum yield = 0.360E+02
Mean elastic modulus = 0.290E+05

No. of Stress Distributions: 2
==========

PRIMARY LOAD FACTOR (# 1)
Entered value for primary: 15.00

S0: Tensile Stress
S0 = 15.00
SECONDARY LOAD FACTOR (# 1)
Entered value for secondary: 36.00

S0: Tensile Stress
S0 = 36.00

PRIMARY LOAD FACTOR (# 2)
Entered value for primary: 5.000

S1: Bending Stress
S1 = 5.000
SECONDARY LOAD FACTOR (# 2)
Entered value for secondary: 0.000

S1: Bending Stress
S1 = 0.000

|--Crack--|----------Crack Sizes----------|---------|----------FAD Op1: Lr vs Kr @ Applied----------
Number a c a1 c1 Bw Bt STATUS 1:Lr 1:Kr(a) 1:Kr(c) 1:Kr(a1) 1:Kr(c1)
|---------|----------|---------|----------
1 0.1000 0.3333 0.1000 0.3333 6.0000 0.3750 DONE 0.4725 0.5624 0.3606 0.5667 0.3606

Abbreviations

ACC	Accuracy
AIC	Akaike information criterion
ANOVA	Analysis of variance
ASME	American Society of Mechanical Engineers
ASNT	American Society for Nondestructive Testing
AWS	American Welding Society
BMCRK	Base metal crack
BSI	British Standards Institute
CASS	Cast austenitic stainless steel
CHL	Coastal and Hydraulics Laboratory
CLCRK	Centerline crack
CNDE	Center for Nondestructive Evaluation
COP	Community of Practice
CRK	Cracks
EDM	Electrical discharge machined
EM	Engineer Manual
ENA	Energy Networks Association
EPRI	Electric Power Research Institute
ERDC	US Army Engineer Research and Development Center
ET	Eddy current testing
FAD	Failure assessment diagram
FDR	False discovery rate

FFS	Fitness for service
FHWA	Federal Highway Administration
FMC	Full matrix capture
FN	False negative
FP	False positive
FST	Fraction of detected flaws sized within tolerances
HSS	Hydraulic steel structure(s)
ID	Identification
IIW	International Institute of Welding
L	Laminar
LAM	Laminations
LOF	Lack of fusion
LOP	Lack of penetration
MT	Magnetic particle testing
NASA	National Aeronautics and Space Administration
NDT	Nondestructive testing
NIST	National Institute of Standards and Technology
NS	Not within tolerances
OECD	Organisation for Economic Co-operation and Development
P	Planar
PAUT	Phased-array ultrasonic testing
PE	Pulse-echo
PISC	Plate Inspection Steering Committee

POD	Probability of detection
POR	Porosity
PP	Total indications
PPV	Positive predictive value
PT	Dye penetrating testing
PVRC	US Pressure Vessel Research Committee
Q-Q	Quantile–quantile
RF	Radio frequency
RMSE	Root-mean-square error
ROC	Receiver operator characteristic
ROCRK	Root crack
RT	Radiography testing
SEN	Sensitivity
SPC	Specificity
SwRI	Southwest Research Institute
T	T-joint
TCG	Time-corrected gain
TFM	Total focus method
TN	True negative
TOCRK	Toe crack
TOFD	Time-of-flight diffraction
TP	True positive
TPR	True positive rate

TRCRK	Transverse crack
UFGS	Unified Facilities Guide Specification
UNK	Unknown
USACE	US Army Corps of Engineers
USAF	US Air Force
UT	Ultrasonic testing
V	Volumetric

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
June 2024	Final Technical Report (TR)	FY19	FY24

4. TITLE AND SUBTITLE

A Study of Phased-Array Ultrasonic Testing (PAUT) for Detecting, Sizing, and Characterizing Flaws in the Welds of Existing Hydraulic Steel Structures (HSS)

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)

Martin T. Schultz, Leslie E. Campbell, Ramsay D. Bell, and Phillip W. Sauser

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

See reverse.

8. PERFORMING ORGANIZATION REPORT NUMBER

ERDC/EL TR-24-12

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
US Army Engineer Research and Development Center Coastal and Hydraulics Laboratory 3909 Halls Ferry Road, Vicksburg, MS 39180-6199	ERDC-CHL	

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES

Funding Account Code U4388268, AMSCO Code 031391

14. ABSTRACT

Hydraulic steel structures (HSS) are components of navigation, flood control, and hydropower projects that control or regulate the flow of water. Damage accumulates in HSS as they are operated over time, and they must be inspected periodically. This is often accomplished using nondestructive testing (NDT) techniques. If damage is detected, the structure's fitness for continued service must be evaluated, which requires information on the location and size of discontinuities. This information can be obtained using ultrasonic testing (UT) techniques. However, there is limited information on the reliability of UT techniques with respect to detecting, sizing, and characterizing flaws in HSS. This study addresses this gap. Round-robin experiments were carried out using phased-array ultrasonic testing (PAUT) to scan weld specimens representing a variety of HSS geometries. The results of the round-robin experiments were analyzed to estimate the probability of detection (POD) and to assess the influence of factors potentially affecting POD. Uncertainty in estimates of flaw length and height were described, and partial safety factors were derived for use in fitness-for-service analyses. These results demonstrate the importance of the technician as a factor influencing the reliability of NDT techniques applied to HSS.

15. SUBJECT TERMS

Hydraulic structures--Maintenance and repair; Nondestructive testing; Phased array antennas; Ultrasonic waves--Testing; Welded steel structures--Maintenance and repair

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	SAR	263

19a. NAME OF RESPONSIBLE PERSON	19b. TELEPHONE NUMBER (include area code)
Martin T. Schultz	(601) 634-4313

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) (concluded)

US Army Engineer Research and Development Center (ERDC)
Environmental Laboratory (EL)
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

US Army Corps of Engineers (USACE)
New Orleans District
7400 Leake Avenue
New Orleans, LA 70118

US Army Corps of Engineers (USACE)
Welding and Metallurgy Technical Center of Expertise
333 SW 1st Avenue
Portland, OR 97204

US Army Corps of Engineers (USACE)
Northwestern Division
1201 NE Lloyd Ste 400
Portland, OR 97232