NUREG/CR-7246
PNNL-27003

U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

Reliability Assessment of Remote Visual Examination

Office of Nuclear Regulatory Research

AVAILABILITY OF REFERENCE MATERIALS IN NRC PUBLICATIONS

NRC Reference Material

As of November 1999, you may electronically access NUREG-series publications and other NRC records at NRC's Library at www.nrc.gov/reading-rm.html. Publicly released records include, to name a few, NUREG-series publications; *Federal Register* notices; applicant, licensee, and vendor documents and correspondence; NRC correspondence and internal memoranda; bulletins and information notices; inspection and investigative reports; licensee event reports; and Commission papers and their attachments.

NRC publications in the NUREG series, NRC regulations, and Title 10, "Energy," in the *Code of Federal Regulations* may also be purchased from one of these two sources.

1. **The Superintendent of Documents**
   U.S. Government Publishing Office
   Washington, DC 20402-0001
   Internet: bookstore.gpo.gov
   Telephone: (202) 512-1800
   Fax: (202) 512-2104

2. **The National Technical Information Service**
   5301 Shawnee Road
   Alexandria, VA 22312-0002
   www.ntis.gov
   1-800-553-6847 or, locally, (703) 605-6000

A single copy of each NRC draft report for comment is available free, to the extent of supply, upon written request as follows:

Address: **U.S. Nuclear Regulatory Commission**
Office of Administration
Multimedia, Graphics, and Storage & Distribution Branch
Washington, DC 20555-0001
E-mail: distribution.resource@nrc.gov
Facsimile: (301) 415-2289

Some publications in the NUREG series that are posted at NRC's Web site address www.nrc.gov/reading-rm/doc-collections/nuregs are updated periodically and may differ from the last printed version. Although references to material found on a Web site bear the date the material was accessed, the material available on the date cited may subsequently be removed from the site.

Non-NRC Reference Material

Documents available from public and special technical libraries include all open literature items, such as books, journal articles, transactions, *Federal Register* notices, Federal and State legislation, and congressional reports. Such documents as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings may be purchased from their sponsoring organization.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at—

The NRC Technical Library
Two White Flint North
11545 Rockville Pike
Rockville, MD 20852-2738

These standards are available in the library for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from—

American National Standards Institute
11 West 42nd Street
New York, NY 10036-8002
www.ansi.org
(212) 642-4900

Legally binding regulatory requirements are stated only in laws; NRC regulations; licenses, including technical specifications; or orders, not in NUREG-series publications. The views expressed in contractorprepared publications in this series are not necessarily those of the NRC.

The NUREG series comprises (1) technical and administrative reports and books prepared by the staff (NUREG–XXXX) or agency contractors (NUREG/CR–XXXX), (2) proceedings of conferences (NUREG/CP–XXXX), (3) reports resulting from international agreements (NUREG/IA–XXXX), (4) brochures (NUREG/BR–XXXX), and (5) compilations of legal decisions and orders of the Commission and Atomic and Safety Licensing Boards and of Directors' decisions under Section 2.206 of NRC's regulations (NUREG–0750).

NUREG/CR-7246
PNNL-27003

U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

Reliability Assessment of Remote Visual Examination

Manuscript Completed: November 2017
Date Published: August 2018

Prepared by:

P. Ramuhalli
P. G. Heasler
T. L. Moran
A. E. Holmes
M. T. Anderson

Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352

Carol A. Nove, NRC Project Manager

Office of Nuclear Regulatory Research

ABSTRACT

Remote visual testing (RVT) is a commonly used nondestructive examination method for inservice inspection (ISI) of reactor internals to detect cracking and gross component failures. Despite widespread use, the detection reliability of RVT and the factors that impact overall RVT performance have been unresolved issues. This report describes the results from an assessment sponsored by the U.S. Nuclear Regulatory Commission and conducted by the Pacific Northwest National Laboratory, in cooperation with the Electric Power Research Institute, for evaluating the reliability of RVT methods currently being used for [reactor] in-vessel visual inspection.

The goal was to assess the performance of commercially applied RVT examination procedures implemented by qualified personnel, as well as to identify and qualitatively assess enhancements to examination procedures for detecting flaws in test specimens. The assessment was performed over three phases of research, with the results of the first phase used to identify key controllable parameters impacting RVT performance and to design the next two phases (round-robin testing). Participants in the round-robin tests included teams of inspectors from commercial nuclear power ISI vendors. Participants were asked to determine if a specimen contained a crack and its approximate location, orientation, and length. The detection and location information were used to compute estimates of the probability of detection and false call probability for various scenarios such as the placement and orientation of the flaw on the specimen, team-to-team variations in detection performance, and effects of secondary review on RVT detection.

Results showed that crack opening displacement (COD) is the dominant factor in the reliability of crack detection using commercially applied RVT procedures, with crack length being weakly correlated with detection probability. The implication is that RVT detection is likely heavily dependent on the contrast produced by the crack opening, with crack detection becoming less reliable as the COD decreases. Further, the results indicated RVT will be challenged when cracks are located in the vicinity of surface features such as scratches or weld ripples, or close to the edge of welds where shadowing and/or the presence of weld undercuts may complicate the ability to detect the crack. The assessment also reinforced earlier findings on the importance of lighting in flaw detection and appeared to align with other studies that find improved reliability when using multiple inspectors or analysts. Finally, the results pointed to the importance of practice with specimens that mimic conditions likely to be found in the field. Based on the findings and limitations of the assessment, a number of recommendations are made regarding best practices for improving the reliability of RVT in a field setting and for addressing remaining informational gaps.

This report describes the design and noted limitations of the three phases of research, the analysis methodology for each phase, and the results of the research. The results of this assessment provide a benchmark set of data on the reliability of RVT for detecting cracking, assuming the implementation of field-like inspection procedures. The likely impact of several uncontrolled factors on RVT detection performance are discussed, and recommendations regarding the use of these results to assess field performance are provided. Finally, recent advances in RVT technology are briefly discussed and point to the potential need for continued research to evaluate the capability and effectiveness of the technique as improvements are implemented.

FOREWORD

Remote visual testing (RVT), an inspection method used extensively by the U.S. nuclear industry to examine reactor components for service-related degradation, is used more often than any other inspection technique for in-service inspections. Its ease of deployment, cost effectiveness, applicability to many different components, and the fact that it is a relatively straightforward and quick process make it a very desirable method to implement. Often, licensees conduct RVT to comply with requirements of the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code (Code), Section XI, *Rules for In-service Inspection of Nuclear Power Plant Components*. However, a significant portion of VT examinations are conducted for inspections of vessel internal components, for which the guidelines are provided in several industry documents such as the BWR VIP-03, *BWR Vessel and Internals Project, Reactor Pressure Vessel and Internals Examination Guidelines* and MRP-228, *Materials Reliability Program: Inspection Standard for PWR Internals*.

In the mid-2000s, the Nuclear Regulatory Commission (NRC) began research on the capabilities and effectiveness of VT to augment the understanding of the method provided by limited studies available in the open literature. NRC funded research at the Pacific Northwest National Laboratory (PNNL) which assessed the important variables that influence the effectiveness of RVT as well as the ability of RVT to detect cracks. Two NUREG/CRs were published as a result of this work, NUREG/CR-6860, *An Assessment of Visual Testing* (2004), and NUREG/CR-6943, *A Study of Remote Visual Methods to Detect Cracking in Reactor Components* (2007). These NUREG/CRs provided the NRC with a greater understanding of the capabilities of RVT and identified many factors that may affect the performance of RVT. However, there were still lingering questions regarding the reliability of RVT methods.

Subsequent to the publication of the two VT NUREG/CRs, the NRC and the Electric Power Research Institute (EPRI) initiated a cooperative round-robin study under the NDE Addendum to the NRC/EPRI Memorandum of Understanding (ADAMS ML16138A556) to assess the capabilities of remote visual testing for detection of surface-breaking cracks. The round-robin study was conducted in three phases:

- Phase I – Preliminary round-robin test to identify key variables that may affect the performance of remote visual examination techniques and develop a test protocol to be used in Phase II.
- Phase II – Round-robin test to quantitatively assess RVT using commercially applied inspection procedures. Statistical analyses such as probability of detection (POD) and false call probability (FCP) were applied to determine the impact that factors such as crack opening displacement, crack length, and surface conditions have on the effectiveness of RVT.
- Phase III – Round-robin test to quantitatively assess the impact to POD and FCP due to enhancements to commercially applied RVT techniques. The assessments included the impact of secondary review of data, and quantify the level of image degradation for recorded data.

This NUREG/CR serves to document the results of the three-phased round-robin study. While the round-robin design and implementation was conducted as a cooperative effort between the NRC/PNNL and EPRI, all data analysis was done independently by PNNL so that findings and conclusions are those of PNNL alone.

The results of this VT round-robin study showed that crack opening displacement is a very important factor in reliably detecting cracks; in other words, as the crack opening gets wider, the resulting difference in contrast produced by the crack opening is critical for reliable detection. The important implication of this finding is that as crack openings decrease, inspection reliability also decreases. Interestingly, while crack length is an important factor used to perform flaw analyses, the length of a crack has less impact on the detectability of a crack with RVT techniques. This work also served to demonstrate that remote visual testing is challenged when cracks are in the vicinity of surface features such as scratches or edges of welds. Additionally, while this study did not delve deeply into the impacts of various lighting modalities, using multiple inspectors and/or independent analysts, and inspector training, the study did provide some insights on these topics.

The results of this research were presented to a joint NRC/Industry public meeting on NDE in January 2017 (ADAMS ML17013A620), and are already being used by industry to revise guidance documents for RVT examinations. This work provides the NRC staff with a precise and statistically significant assessment of the types of flaws that one can expect remote visual testing to be able to find in the field. The NRC staff will use this work to inform a Code Case Regulatory Guide rulemaking on visual testing-related issues, such as the use of visual testing in lieu of ultrasonic testing for nozzle inner-radius examinations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This report describes the results from an assessment sponsored by the U.S. Nuclear Regulatory Commission (NRC) and conducted by the Pacific Northwest National Laboratory (PNNL), in cooperation with the Electric Power Research Institute (EPRI), for evaluating the reliability of remote visual testing methods currently being used for [reactor] in-vessel visual inspection (IVVI).

Remote visual examination or remote visual testing (RVT) is a commonly used nondestructive examination (NDE) method for inservice inspection (ISI) of reactor internals to detect cracking and gross component failures. RVT for crack detection uses cameras and other equipment for inspecting components in hard-to-access regions, with acceptable practices described in several industry-controlled documents.

This work was originally motivated by industry interest in pursuing, through the American Society of Mechanical Engineers (ASME) Code, supplanting volumetric examinations with visual examination. However, there were open questions on the performance of RVT, especially in terms of probability of detection (POD), relative to other NDE methods.

This report describes results of the assessment, conducted over three phases, for answering the open questions. The goal was to assess the performance of commercially applied examination procedures implemented by qualified personnel. Specific objectives were:

- Identify factors that may influence the POD of RVT, and quantify the reliability of RVT procedures as a function of these factors.
- Identify and qualitatively assess enhancements to commercially applied procedures for RVT for detecting flaws in test specimens.

Given the range of possible contributors to RVT effectiveness and the difficulty in conducting controlled assessments that address every factor in a reasonable amount of time, PNNL and EPRI made a conscious decision to limit the scope of the study to assessing the impact of key variables on the ability to detect (but not necessarily measure) cracks. This included variables related to the size of the indications, including length and crack opening displacement (COD), and the effect of factors that may confound the ability to reliably detect indications. Other factors, such as the inspection speed, were bounded in accordance with ASME Code requirements, while instrumentation factors (camera resolution, lighting, etc.) were largely uncontrolled. Finally, factors such as oxidation of the specimens, the turbidity in the water, thermal effects, irradiation degradation of camera imaging capability, and convective flow effects are all known to affect the ability to detect cracking using RVT; these factors were designed out of the study and were not addressed. As a result of these choices, the study provides a benchmark set of data on the reliability of RVT (specifically, VT-1 and enhanced visual testing (EVT-1)) for detecting cracking, assuming the implementation of field-like inspection procedures, but without uncontrolled factors that may affect (positively or negatively) the detection performance.

Phase I of this assessment was used to identify key controllable parameters impacting RVT performance. These results were used to design and conduct a round-robin exercise (Phase II). This was followed in Phase III by an assessment, again using a round-robin exercise, of potential improvements in RVT procedures for IVVI to enhance the ability to discriminate between cracks and non-relevant indications such as surface features.

Participants in the round-robin tests included teams of inspectors from commercial nuclear power ISI vendors. Each of the tests was “blind” in that the true condition of the specimens used was not revealed to the test takers. Participating teams were asked to determine if a specimen contained a crack and its approximate location, orientation, and length. The detection and location information were used to compute estimates of the POD and false call probability for various scenarios such as the location of the flaw on the specimen, team-to-team variations in detection performance, and effects of secondary review on RVT detection.

Results showed that COD is a major factor in the reliability of crack detection using commercially applied RVT procedures, with crack length being weakly correlated with detection probability. While smaller COD values are usually associated with shorter cracks, limited operational experience has shown that some forms of degradation, such as stress corrosion cracking (SCC), can result in a small COD for longer cracks. From a practical standpoint, the finding from this assessment implies that RVT detection is likely heavily dependent on the contrast produced by the crack opening, with crack detection becoming less reliable as the COD decreases. Note that unreliable detection is not the same as no detection. It simply means the probability that the crack will be detected every time is low. Further, RVT will be challenged when cracks are located in the vicinity of surface features such as scratches or weld ripples, or close to the toe of welds where shadowing and/or the presence of weld undercuts may complicate the ability to detect the crack. These hypotheses were supported by the results from Phases II and III and point to some limitations of RVT.

The assessment reinforced earlier findings regarding the importance of lighting in RVT flaw detection and appeared to align with other studies that find improved reliability when using multiple inspectors or analysts. The results also pointed to the importance of practice, especially with specimens that mimic the conditions likely to be found in the field.

Extrapolating the results to a typical field inspection is a challenge and will require augmenting data from this assessment with quantities such as the surface feature density in field components and false call rates (FCRs) in the field. In addition, the effect that camera deployment systems (by rope/pole or robotic devices) will have on the detection performance is unknown. Given these open questions, some of which would be extremely challenging to quantify, the results should be considered a baseline and a means to help identify factors that may influence the POD and FCR in a field examination. Results indicate that improvements in POD can be obtained through specific actions, such as better training; therefore, industry actions should be considered for implementation with the expectation they will lead to improvements in field inspection performance.

Required detection and FCRs for field inspections are also unclear as these need information (unresolved in this assessment) on acceptance criteria for flaws. The acceptance criteria information is often plant- and component-specific. The findings from this assessment can help identify factors that may limit the ability to detect critical flaws in specific plants and components. Examples of such factors are the presence of surface features, crack adjacency to the weld toe, or other geometric/physical features that impact optimizing applied lighting and viewing angles. Understanding the impact of these factors on detection can lead to potential improvements in RVT instrumentation as well as the recommendation to use other inspection methods, such as ultrasonic testing, that may provide the necessary sensitivity for detecting cracks in conditions and components unfavorable for RVT.

As stated earlier, several variables may affect RVT performance but were intentionally left out of these studies. These include oxide buildup on internal components; thermal distortion of video images; water currents and clarity; radiation effects on camera video quality; limits on accessibility, viewing angle, and lighting; camera delivery systems; and personnel qualification levels. While all of these factors are expected to impact RVT performance to different levels, we expect that the impact of some of these factors may be limited by procedures used in the field.

Based on the findings and limitations of the assessment, the following recommendations are made (in no particular order of importance):

- RVT procedures should be updated to include additional details on performing the inspection and guidance for discriminating between cracks and non-cracks. While this information may be ingrained in the knowledge base of experienced analysts, such information may be helpful as a reminder for all analysts.
- Representative specimens that mimic the surface conditions and types of cracks likely to be encountered in the field should be used for training purposes prior to inspection teams performing field examinations. While prerecorded video data may be applied for this purpose, they are unlikely to provide sufficient opportunities for exercising skills in discriminating between cracks and non-cracks.
- The limitations of RVT should be in the forefront when planning or analyzing data from an inspection. Consideration should be given to the use of alternate techniques such as ultrasonic testing and eddy current testing for inspecting challenging areas such as weld toe regions.
- The applicability of RVT should be determined in close conjunction with the development of crack acceptance criteria specific to the components being inspected. In many cases, it is likely that large cracks can be tolerated, such as in the case of core shrouds in boiling water reactors (BWRs). In these cases, the reliability of RVT should be sufficient to detect cracks well before failure of the component. In other instances, where much smaller cracks need to be detected, the specific circumstances associated with the component need to be considered prior to the application of RVT; for example, the environment, minimum detectable flaw size, and the impact of missed detections.
- Camera deployment systems used will likely affect the overall reliability of the examination. This effect needs to be better quantified.
- Advances in RVT technology, such as high-definition cameras and automated image analysis algorithms, should be evaluated to determine if these can help further improve the reliability of RVT.
- The condition of the surface (texture, patina, oxide or other deposits) may be important in detection; however, this assessment did not extensively evaluate these factors, nor did it evaluate the effects of these deposits and the effectiveness of cleaning procedures. The impact of these factors needs to be better quantified.
- While a review of the detection results by a secondary analyst appeared to be effective at reducing false calls, this assessment was not designed to thoroughly evaluate the possible benefits of teams of inspectors or analysts. A further evaluation of these factors using well-controlled and well-designed human-factors studies will be needed for better quantification of the benefits of inspection teaming efforts, if these data are deemed important.

Potential uses for visual examinations in nuclear power applications seem to be increasing, with proposed use of these methods to inspect spent nuclear fuel dry storage canisters and for advanced reactors, such as liquid metal and high-temperature gas reactors. As small modular reactors come on line, it is expected that VT in general, and RVT in particular, may also play a role in assuring the integrity of components. However, these newer applications appear to bring additional challenges with respect to types and location of cracking, access restrictions, and cracking precursors that may challenge existing instrumentation and procedures, and may require additional skills development for RVT inspection teams. Proposed automated analysis techniques for RVT are also likely to become commonplace. Such techniques were discussed during the development and conduct of this assessment, but were ultimately not included in the tests as the technology was not deemed to be sufficiently mature.

These developments in RVT technology and anticipated challenges in applying VT to different systems point to the need for continued evaluation of the capability and effectiveness of this inspection technique. Given these developments, it is also likely that there may be a renewed push to use VT over other NDE techniques, and VT may be the only option in some cases. As a result, it may be appropriate for future work to include studies that benchmark the performance of VT with respect to other NDE methods targeted for specific components and cracking mechanisms.

ABBREVIATIONS AND ACRONYMS

A	axial
ASME	American Society of Mechanical Engineers
B&W	black and white
BPV	Boiler and Pressure Vessel Code
BWR	boiling water reactor
BWRVIP	BWR Vessel and Internals Project
C	circumferential
COD	crack opening displacement
DOF	degrees of freedom
EP	examination procedure
EPRI	Electric Power Research Institute
ET	eddy current testing
EVT-1	enhanced visual testing-1
FCP	false call probability
FCR	false call rate
G	ground (as in weld)
GLM	general linear model
GOF	goodness-of-fit
GrWeld	ground weld
GU	grading unit
H	horizontal
HAZ	heat-affected zone
HD	high-definition
ID	inner/inside diameter or identification
IGSCC	intergranular stress corrosion cracking
InSF	in surface feature
ISI	inservice inspection
IVVI	in-vessel visual inspection
MRP	Materials Reliability Program
MT	magnetic particle testing
N	no
NDE	nondestructive examination
NG	not ground
NGW	not ground weld
NOBS	number of observations
NRC	U.S. Nuclear Regulatory Commission

OD	outer/outside diameter
PNNL	Pacific Northwest National Laboratory
POD	probability of detection
PODF	final POD as determined by secondary analyst
PODP	POD as determined by primary analyst
POR	probability of recording an indication by primary analyst
PT	penetrant testing
PWR	pressurized water reactor
R	review
RES	Office of Nuclear Regulatory Research
RMS	root mean square
RMSE	root mean square error
ROC	receiver operating characteristic
RPV	reactor pressure vessel
RRT	round-robin test
RT	radiographic testing
RV-RRT	Phase II round-robin test
RV-RRT-3	Remote Visual Round-Robin Test Phase III
RVT	remote visual testing
S	flaw size
SCC	stress corrosion cracking
SF	surface feature (scratched area)
SS	stainless steel
TGSCC	transgranular stress corrosion cracking
U.S.	United States
UT	ultrasonic testing
V	vertical
VT	visual testing
Y	yes

Definitions

This glossary defines some important terms used in the report.

Blank Grading Unit/Material	A unit of material that contains no flaw. In this study, a blank grading unit may either be clean (no scratches or other benign surface features) or include benign surface features (such as scratches and grind marks but not cracks)
FCP	False call probability. This is the probability that a blank unit of material is called defective.
FCR	False call rate. This is the number of false calls per unit length of inspected weld.
POD	Probability that a unit of material is called defective. POD(X) represents probability of detection calculated under condition X.

ACKNOWLEDGMENTS

The work described in this report was sponsored by the Office of Nuclear Regulatory Research (RES) at the U.S. Nuclear Regulatory Commission (NRC). The authors gratefully acknowledge the guidance provided by the NRC Contracting Officer Representatives, Mr. Wallace E. Norris (RES, Retired) and Ms. Carol Nove (RES) during the course of this study. Dr. Stephen Cumblidge at the NRC Office of Nuclear Reactor Regulation (NRR) also provided insights into remote visual testing and the NRC's interest. We are also grateful to the Electric Power Research Institute (EPRI), and in particular, to Mr. Jeffrey Landrum, Mr. Chris Joffe, and Mr. John Lindberg at EPRI for their active participation in this coordinated research activity between the NRC and EPRI. Their assistance in helping understand the current state of remote visual inspection in the nuclear power industry, and background information on current standards, was extremely valuable in placing the results of this work in context. The support provided by industry participants in the round-robin testing exercise was also invaluable in the performance of this research. Finally, we gratefully acknowledge Ms. Kay Hass and Ms. Janice Haigh for their invaluable assistance in the technical editing and formatting of this report, and their patience during this process. The authors also thank the technical peer reviewers for their feedback and assistance in improving this report.

1 INTRODUCTION

Remote visual examination or remote visual testing (RVT) is a commonly used nondestructive examination (NDE) method for inservice inspection (ISI) of reactor internals to detect cracking and gross component failures. RVT for crack detection is considered an enhanced version of the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel (BPV) Code (ASME "Code") VT-1 examination. Acceptable practices for enhanced VT-1 (so-called EVT-1; EPRI 2005; Landrum and Selby 2005) are described in several industry-controlled documents such as BWR Vessel and Internals Project (BWRVIP)-03 (EPRI 2005) and Materials Reliability Program (MRP) MRP-227/MRP-228 (EPRI 2015a, b).

Visual examination, and consequently RVT, is considered to be a relatively straightforward method for inspection of components, with little skill (IAEA 2013) assumed to be needed for successful inspection. In comparison with other volumetric and surface examinations, such as Code-approved ultrasonic and eddy current examinations, RVT examinations can be performed faster and with potentially less radiation exposure to personnel. The cost to deploy equipment and inspection personnel for visual inspection is also generally considered to be low when compared to the cost of other NDE methods. Given these potential benefits, it is not surprising that ASME Code Cases provide for the use of VT-1 examinations in lieu of ultrasonic testing (UT) examinations for inner radius inspections of Class 1 reactor vessel nozzles (ASME 2010) and Class 1 pressurizer and steam generator nozzles (ASME 2015d). These Code Cases have been determined by the U.S. Nuclear Regulatory Commission (NRC) as conditionally acceptable alternatives (NRC 2014) to applicable parts of ASME Code Section XI (ASME 2015e).

A major open question with regard to the use of RVT in lieu of other volumetric or surface examination methods is how the reliability of RVT compares with that of volumetric examination methods for the purpose of detecting relevant cracking. A key aspect to addressing this question is the determination of the baseline reliability of remote visual examinations. This report describes the results of an assessment conducted to determine the reliability of RVT. These results specify a process for quantifying the performance of RVT for ISI purposes and provide a baseline for RVT performance under controlled conditions.

The NRC sponsored Pacific Northwest National Laboratory (PNNL) to conduct this study for evaluating the reliability of RVT methods currently being used for [reactor] in-vessel visual inspection (IVVI). The work was done cooperatively with the Electric Power Research Institute (EPRI) NDE Center under the NDE Addendum (NRC 2011) to the NRC/EPRI Memorandum of Understanding. A phased approach was used in the assessment, with a limited, preliminary round-robin test (RRT) conducted during Phase I of this evaluation. The results of the Phase I test, along with a subsequent parametric study, were used to design and conduct a more extensive Phase II round-robin exercise (designated as the RV-RRT in this document). Phase II results pointed to the possibility of improving procedures for IVVI to enhance the ability to discriminate between cracks and non-relevant indications such as surface features (SF). These results, along with other parametric work, were used to design and conduct a Phase III round-robin activity, henceforth designated as the RV-RRT-3.

1.1 Assessment Objectives

This report documents the approach and results from the NRC-EPRI coordinated research activities for evaluating the reliability of RVT. The goal was to assess the performance of commercially applied examination procedures implemented by qualified personnel. Within the context of this study, qualification refers to the minimum requirements for certification of NDE personnel (ASNT 2016a, b).

Specific objectives of this study were:

- Identify factors that may influence the probability of detection (POD) of RVT and quantify the reliability of RVT procedures as a function of these factors.
- Identify and qualitatively assess enhancements to commercially applied procedures for RVT for detecting flaws in test specimens.

Several variables are likely to impact the reliability of RVT, including material, inspection and equipment parameters, and human factors.

The study described in this report focused on certain material, environmental, and equipment parameters only, and deferred the quantification of the effects of human factors on RVT reliability to future studies. Consequently, the results described in this report should only be considered as a baseline for the performance of RVT and may not fully reflect performance in field conditions.

1.2 Report Outline

Section 2 of this report contains a description of RVT, including an overview of equipment, procedures, and applicable codes and standards. Section 3 provides an overview of the phased approach used for quantifying performance of RVT. Sections 4 through 6 describe in greater detail the experimental design, analysis results, and key findings from each of the three phases of the study. Section 7 discusses the overall findings and places these findings in the context of IVVI and other ISI requirements in the nuclear power industry. Finally, Section 8 draws conclusions from the results presented, and briefly outlines recommendations.

A set of appendices is also included, with Appendix A and Appendix B listing the protocols for conducting the round-robin studies in Phase II and Phase III, respectively. Appendix C describes the procedures used for determining the true-state information about crack opening displacement (COD) and length, while Appendix D provides a tutorial on POD analysis. Appendix E and Appendix F provide a more detailed description of the results from Phases II and III, respectively. Appendix G describes an assessment of different regression models for their applicability to the Phase III data.

2 BACKGROUND

Prior PNNL studies (Cumblidge et al. 2004; Cumblidge et al. 2007) have examined RVT for nuclear power applications. These studies, which were all parametric in design, have generally found that a number of factors may affect the application and performance of RVT. Details of these parametric studies are available in the cited reports; this report summarizes the state of the art in RVT for the nuclear power industry and briefly describes new technology that may improve RVT.

2.1 Remote Visual Testing

RVT uses cameras and other equipment for inspecting components in hard-to-access regions. RVT in general follows the ASME Code specifications for performing visual tests. Three classes of visual examination are described in the ASME Code—VT-1 for crack detection and sizing, VT-2 for leak detection, and VT-3 for assessing the gross mechanical condition of components. Requirements for inspection and personnel training and the acceptance criteria are different for each of these classes of VT.

RVT uses a number of aids for performing the inspection given the requirement for the inspector to be at a distance from the component. Camera systems (including borescopes) are typically used to view areas of the component or plant that have limited accessibility. For in-vessel inspection, the cameras are generally radiation-hardened. Both color and black/white cameras are currently in use, and many cameras provide pan-tilt-zoom capability along with on-camera lights for improving image quality. More recently, manufacturers have started making available high-definition (HD), radiation-tolerant cameras for use in IVVI, although these cameras have not yet seen widespread use.

In addition to the on-camera lights, auxiliary lighting, particularly diffuse lighting (Cumblidge et al. 2007), may be used to enhance a crack or help differentiate between cracks and other surface features. These lights are usually deployed separately from the camera system. For in-vessel inspection, the deployment options are currently limited and usually based on a rope/pole arrangement that is held in position by the camera operator(s). Such systems are typically subject to camera wobble that is believed to affect the ability to detect cracks. While alternative systems such as robotic manipulators have been proposed and are under development, these require further evolution to ensure engineering and operational maturity before becoming available for use.

Note that the deployment of cameras into hard-to-access regions is not limited to in-vessel inspections. For spent nuclear fuel dry storage cask inspections, for example, borescopes or other cameras will need to be delivered to the annulus between the cask and overpack through narrow access channels. Again, delivery of these systems to the desired location is challenging and may require advanced robotics and considerable skill on the part of the inspector.

Data from these cameras are usually streamed to a recording device, typically a desktop computer with standard video recording equipment. Audio tracks are often included and capture notes or other commentary from the inspector. The amount of video recorded is large (several gigabytes), necessitating use of video compression algorithms. Experience has shown that most camera systems use standard compression algorithms (such as MPEG-4) for this purpose. Note that some systems will also allow the inspector to obtain a snapshot image of specific regions for documentation purposes.

The use of recorded video and still images brings up the intriguing possibility that automated analysis software could be applied to find and characterize cracks. Such techniques are described in the literature (for instance, Newman and Jain 1995; Pascual 2014; Schmugge et al. 2014; Liu et al. 2015; Chen et al. 2017). To our knowledge however, automated video analysis technology for crack detection and characterization remains a research and development area that is being actively pursued.

2.1.1 Overview of Procedures

PNNL and EPRI reviewed a number of procedures by participants in the RRTs conducted for this assessment; however, these procedures were made available under individual nondisclosure agreements with the participants and are not included in this report. A brief overview of typical attributes for RVT procedures based on general experience and discussions with EPRI and other participating organizations is included below.

Most commercially applied procedures for remote visual examination include:

- Applicability restrictions that describe the components and environments within which the procedure may be applied
- Equipment that may be used for the inspection
- Calibration procedures, which typically involve a resolution check where the standard (usually the ASME Character Standard) is placed at one or more fixed distances from the camera and the operator confirms the visibility of the characters
- Pre-inspection planning and setup, including any required cleaning of the component and resolution check (cleaning procedures are generally used in cases where oxide buildup on the surface being examined may mask the presence of cracks)
- General guidance on inspections and interpretation of data, which include directions for ensuring that the inspection coverage is adequate and information on the types of indications that may be encountered (including non-crack indications such as surface features that may be present)
- Reporting guidance.

2.1.2 Applicable Codes and Standards for RVT

Within the nuclear power community, the following Codes and industry guidance are typically used for inspections. Note the guidance cited is specific to in-vessel (i.e., internal) components. A brief summary of these guidelines is provided below.

- ASME VT-1: ASME Code, Section XI, and Section V, Article 9 provide guidance on visual testing (VT-1) and how to substitute remote visual examination for direct examinations. The remote examination procedure shall demonstrate the same capabilities as direct visual. Additionally, the remote examination system needs the capability to distinguish and differentiate between the colors (such as different shades of gray for steels) applicable to the component examination being conducted (IWA-2211 (g), ASME 2015a).
- BWRVIP-03: The BWR Vessel and Internals Project (BWRVIP-03) contains guidelines for managing degradation in reactor vessel internal components, vessel welds, and nozzles. Among the various activities overseen by BWRVIP is the development of NDE techniques for assessing the integrity of these components. This inspection guidance supplements mandated

examinations, and includes generic standards for visual examination of boiling water reactor (BWR) internal components. Remote visual examinations using a high-resolution camera, with the potential to magnify the image if the camera includes this capability, is referred to as an enhanced visual examination (EVT-1) (Landrum and Selby 2005). The enhancements are intended to improve the detection and characterization of discontinuities during reactor internals examinations.

- MRP-227: The Materials Reliability Program (MRP) 227, *Pressurized Water Reactor Internals Inspection and Evaluation Guidelines*, contains guidelines for managing long-term aging in reactor vessel internal components of pressurized water reactors (PWRs) (EPRI 2015a). In MRP 227, EVT-1 is recommended for remote visual examinations.

2.2 Motivation for RVT Reliability Assessment

This work was originally motivated by industry interest in pursuing, through the ASME BPV Code, supplanting volumetric examinations with visual examination. In comparison with volumetric and surface examinations, RVT examinations can be performed faster and with less personnel exposure to radiation. Additionally, RVT examiners are generally subjected to a less rigorous qualification process when compared to other NDE methods. However, there were open questions on the performance of RVT, especially in terms of POD, relative to other NDE methods.

Code Cases, such as N-619 and N-648-1, have been published in recent years that allow for the use of a VT-1 examination in lieu of a UT examination (ASME 2010, 2015d) of nozzle inner radius of Class 1 pressurizers, steam generators, and reactor vessel nozzles, and have been conditionally accepted by the NRC (NRC 2014). Specifically, licensees are allowed to perform a VT-1 examination in lieu of a UT examination, with the limiting surface flaw length size calculated using the allowable surface flaw sizes from Table IWB-3512-1 (ASME 2017) and a limiting flaw aspect ratio of 0.5 (78 FR 37885). However, concerns were expressed with the ability for RVT to reliably detect smaller service-induced cracks in these components (78 FR 37885).

VT as an inspection method continues to be widely employed and is expected to see more applications as existing reactors age. Increased RVT use is also anticipated as new requirements for inspection arise, such as for inspecting dry storage casks for atmospheric stress corrosion cracking (SCC), and as new reactor technologies (e.g., small modular reactors and fast or high-temperature reactors) become licensed to operate.

Whether in the context of RVT in lieu of volumetric examination methods, or for other anticipated applications, several issues need to be addressed. These include:

- Identification of key variables that may impact the performance of RVT.
- Determination of the performance of RVT, in terms of the POD and false call probability (FCP) as a function of the key variables described above. This information will provide insights into the smallest flaws that can be detected reliably by procedures and systems, as well as information on improvements to equipment and procedures that may increase the overall performance of RVT for IVVI and other uses.
- Determination of acceptance criteria for RVT in terms of the largest flaws that can be tolerated. Note that this may be plant- and component-specific. NUREG/CR-6943 (Cumblidge et al. 2007) discussed a few examples of acceptance criteria and noted the variability by plant and component.

- Determination of a basic methodology for comparing different NDE methods for the specific purpose of using them in lieu of other inspection methods, such as volumetric vs. surface or surface vs. surface. Fortunately, such a methodology has been described elsewhere (Forli 1995; Moran et al. 2010) and can be applied to the present question.
- Quantification of performance data for other (non-RVT) inspection methods.

This assessment focuses on the first two issues listed above, namely RVT key variables and performance, including COD and crack length measurements that are common to specific degradation mechanisms such as fatigue cracks, intergranular stress corrosion cracking (IGSCC), transgranular stress corrosion cracking (TGSCC), pitting, and other similar degradation mechanisms; and minimum CODs that are detectable using enhanced magnification remote VT-1 examinations. Depending on the answers to these questions, additional research may be needed to address remaining issues such as detection of cracks in the presence of certain conditions (e.g., oxides, dirt, scale, or other component surface features).

3 OVERVIEW OF APPROACH TO REMOTE VISUAL EXAMINATION PERFORMANCE ASSESSMENT

Previous research (Spencer 1996; Cumblidge et al. 2007) identified several factors that are likely to impact the performance of RVT, especially remote visual examination conducted in accordance with VT-1 and EVT-1. These include, but are not limited to, visual acuity of the system, size of indications including the length and COD of the indication, contrast between indication and surface, scanning or inspection speed, surface conditions, light levels, lighting angle, camera angle, and human factors.

Given the range of possible contributors to RVT effectiveness and the difficulty in conducting controlled studies that address every factor in a reasonable amount of time, PNNL and EPRI made a conscious decision to limit scope to assessing the impact of key variables on the ability to detect cracks, but not necessarily measure their size. However, the included variables were directly related to the size of the indications and the effect of factors that may confound the ability to reliably detect indications. These factors included surface condition, presence of surface features, and location of cracks. Other factors, such as the inspection speed, were bounded in accordance with ASME Code requirements, while instrumentation factors (camera resolution, lighting, etc.) were largely uncontrolled. Finally, factors such as oxidation of the specimens, turbidity in the water, thermal effects, irradiation degradation of camera imaging capability, and convective flow effects are all known to affect the ability to detect cracking using RVT; however, for simplicity and efficiency these factors were purposely not included in the study.

As a result of these choices, the assessment provides a baseline set of data on the reliability of RVT for crack detection, assuming field-like inspection procedures but without uncontrolled factors that may positively or negatively affect the detection performance.

The large number of influencing factors and the potential for interactions between multiple factors led to the choice of a phased approach for this assessment, with limited RRT conducted during Phase I. The results of Phase I were used to identify key factors that may play a large role in the detection performance of RVT, and to design and conduct Phase II for a more thorough evaluation of these factors. The results of Phase II, along with a subsequent parametric study, were used to design and conduct the Phase III RRT (designated as the RV-RRT-3 in this document). Details of the objectives and expected outcomes of each phase are provided below.

3.1 Phase I

Phase I included the following objectives:

- Identify key variables that may affect the performance of remote visual examination techniques for detecting and characterizing flaws in test specimens.
- Identify test protocol options for Phase II RRT.

Specific variables evaluated during Phase I included COD and length, typical camera systems used for RVT, specimen type and surface conditions, and a feasibility assessment of the methodology for formal RRT.

A mini-RRT, followed by a set of parametric studies, was implemented to meet the objectives. The results of Phase I were used to design and conduct a more extensive Phase II RRT (designated as the RV-RRT in this document). This process included fabrication of multiple specimens for Phase II and development of an appropriate test protocol.

3.2 Phase II

The Phase II RV-RRT had the following objectives:

- Identify and quantitatively assess remote visual examination techniques for detecting and characterizing flaws in test specimens.
  - Evaluate commercially applied inspection procedures for their effectiveness
    - Quantify procedure performance in terms of POD and determine the effect that certain important factors have on POD. Important factors for Phase II include:
      - Crack opening displacement
      - Crack length
      - Crack detection in the presence of surface irregularities or blemishes.

The RV-RRT data from Phase II were expected to provide a better overall understanding of the performance of commercially applied remote visual examination procedures and the critical factors that affect performance.

In addition, the data were expected to be sufficient to calculate:

- POD curves for each participating inspection team as a function of flaw size (COD and length)
- Identification of significant differences in POD related to important variables, including examination procedure, flaw type, orientation, and flaw location. A limited assessment of the effect of surface patina was included.
- Evaluation of FCP.

3.3 Phase III

Phase III of this study (designated as RV-RRT-3) extended the Phase II objectives and added the following:

- Identify and quantitatively assess enhancements to remote visual examination techniques for detecting and characterizing flaws in test specimens
  - Evaluate improvements to commercially applied examination procedures for their effectiveness
- Quantify the impact of secondary review of all recorded data in terms of changes in flaw detection rates and false call rates (FCRs)
- Quantify the level of image degradation (if any) in recorded data
- Quantify procedure performance in terms of POD and determine the effect that certain important factors have on POD (important factors for Phase III were the same as Phase II):
  - Crack opening displacement

- Crack length
- Crack detection in the presence of surface irregularities or blemishes.

In the remainder of this report, we describe the experimental design for each of these phases and the key results obtained from each phase. While alternate ways of organizing the information are possible, we describe the information sequentially (Phase I, followed by Phases II and III), as the findings from each phase motivated subsequent work and informed the design of the experimental study for each phase.

Overall, the research from each phase reinforced the findings from the previous phase and led to a new understanding of the capabilities and limitations of RVT. Key overall findings are summarized after a description of Phase III, in Section 7.

3.4 Limitations and Assumptions

Several factors that impact field RVT were not included in these assessments. The primary reasons for the exclusion were a perceived inability to quantify the magnitude of a particular issue in a simple manner, and to arrive at findings more expeditiously. Factors not evaluated include:

- Surface patina, and oxide or "crud" buildup on internal components. While an oxide presence may influence crack detection negatively by masking the crack or positively by "decorating" the crack, most field examination procedures require cleaning when oxide buildup is excessive. This is an issue with primarily BWR plants and believed not likely to be a major factor for PWRs. Given these considerations, this factor was not included in this evaluation. Effectiveness of cleaning procedures was also not evaluated, given the difficulty in creating realistic and repeatable oxide buildup for testing of cleaning procedures. Instead, the specimens were fabricated using oxide-resistant materials (stainless steel and ceramic). Ceramic specimens were uniformly colored to simulate a generally encountered surface patina within BWRs. Stainless steel (SS) specimens in Phases II and III were subjected to a media-blast to create a matte-like surface that reduces specular reflectivity and mimics typical surface patinas in internal components in PWRs and some BWRs. All specimens were only used in a controlled immersion environment and thoroughly cleaned and dried after use to limit the formation of any deposits on the specimen surface. Additional details on specimen design and fabrication are provided in Sections 4 through 6.
- Thermal distortion of video images, water currents, and water clarity. While each of these factors may affect the video quality from RVT cameras, conducting a viable test requires the ability to quantify these factors. Given the typical standoff distances used in RVT (generally less than 12 inches), these factors may not be very influential. As with the oxide buildup issue, quantifying the magnitude of these factors is necessary before they can be formally included in a test.
- Radiation effects on camera video quality. Currently, degradation of video quality is monitored and cameras are replaced as image quality becomes an issue. It is unclear if this assessment is subjective or if objective metrics of video quality are used. In either case, given that monitoring of image quality for radiation degradation effects is done continuously, this factor was not considered significant for this study.
- Camera delivery systems. Currently, most RVT inspections in the field are performed using a rope/pole system for camera delivery. Experienced teams have been known to deliver the camera and perform the inspection with minimal jitter of the camera/delivery platform. Anecdotal evidence also indicates that industry is considering robotic delivery systems for

many locations, which may reduce the need for further evaluating this specific factor. However, such systems without the ability to precisely deliver and control the camera may negatively impact the ability to detect cracks. Given that this effect is likely to be a function of the specific component, its location within the reactor vessel, the critical flaw size for the component, and other issues related to personnel skills for camera maneuverability, this effect is not included in this study. It is likely that a parametric study may be sufficient to address this question, along with human factor considerations, if necessary.

- Physical access, angle of view limits, and specimen geometry (radius of curvature). These three factors are related and will affect distance of the camera to the component, impacting field of view and angle of view. These were not addressed given the timelines involved in conducting the assessment. Instead, all specimens used in this assessment were flat and the test protocols allowed access to all parts of the specimen.
- Personnel qualification and training. These factors were not formally controlled and, as discussed in Sections 4 and 6, data are suggestive but not conclusive. Findings such as the value of practicing inspection with representative specimens, supplementary guidance for evaluating indications found during RVT examination, and the effect of teaming on performance should be considered only a first step towards addressing significant human factors elements in RVT, with follow-on studies needed if these do not appear to be adequate. Evaluating the effect of these factors will require a larger, well-designed human factors assessment.

Critical flaw sizes for reactor internals are a necessary quantity for assessing the overall capability and reliability of RVT. Such fundamental data form the baseline for the smallest flaw that must be reliably detected and provide the context for evaluation of the POD data derived from studies such as reported here.

Determining the critical flaw sizes for reactor internals has proven to be a challenge, primarily due to the diversity in materials and component dimensions as well as the stresses endured by the different components. Such calculations, typically based on fracture mechanics assessments, must necessarily be performed in a plant-specific manner, though such calculations tend to be proprietary. Previous studies (Cumblidge et al. 2007) provided examples of acceptable flaws for various components, calculated under several assumptions, and indicated that many components may be able to tolerate relatively long flaws with no impact on operability or plant safety. However, it is unclear how representative these calculations are.

Given the lack of such information, the present study was limited to postulating a range of flaw dimensions and evaluating the POD of typical RVT procedures. The resulting data provide information on flaw dimensions (length and COD), above which reliable detection—defined in this report as 80% POD at a 95% confidence level—is generally possible. This does not imply that smaller flaws are not detectable; it simply states that under the conditions used in this assessment, flaws smaller than the reliable detection threshold were difficult to detect with the procedures used. As a result, it is possible that the resulting data may constitute an upper bound on performance relative to detection of small flaws under field conditions.

Also beyond the scope of the present study was a comparison of RVT against other commonly used volumetric and surface examination techniques. Additional studies would be needed to quantify the effectiveness of EVT-1 and VT-1 relative to the effectiveness of UT performed from the outside diameter (OD) to detect flaws on the inside diameter (ID), UT performed on the ID to detect flaws on the ID, and other NDE methods such as radiographic testing (RT), penetrant testing (PT), magnetic particle testing (MT), and eddy current testing (ET).

4 PHASE I OVERVIEW AND RESULTS

Phase I, as described in Section 3, was intended to identify key variables and RRT protocol options for conducting a formal set of tests towards RVT performance assessment.

4.1 Experimental Design

Phase I was conducted as a mini-RRT, followed by a set of parametric studies. Key design variables were crack COD, orientation relative to the observer, and specimen surface finish.

4.1.1 Specimen Design

Two types of specimens were used and are summarized in Table 4-1.

Table 4-1 Summary of Specimen Design for Phase I

Type of Test	Specimen Shape and Size	Material	Surface Condition
Mini-Round Robin	Round (~50 mm [2 in.] diameter)	Stainless steel	As-received
Parametric Study	Rectangular (~50 × 150 mm [2 × 6 in.])	Stainless steel	Matte finish
Parametric Study	Rectangular (~280 × 280 mm [11 × 11 in.])	Ceramic	As-received

The specimens used for the mini-RRT were designed to identify variables that may affect the performance of remote visual examination techniques for detecting flaws in test specimens. The specimens used for the mini-RRT had a fiducial mark indicating the 0° location (Figure 4-1). Specimens could be rotated so that the fiducial was located at 0°, 90°, 180°, or 270° from the vertical. This allowed the physical location and orientation of the cracks to vary relative to the observer and tested the effect of flaw orientation on detection performance.

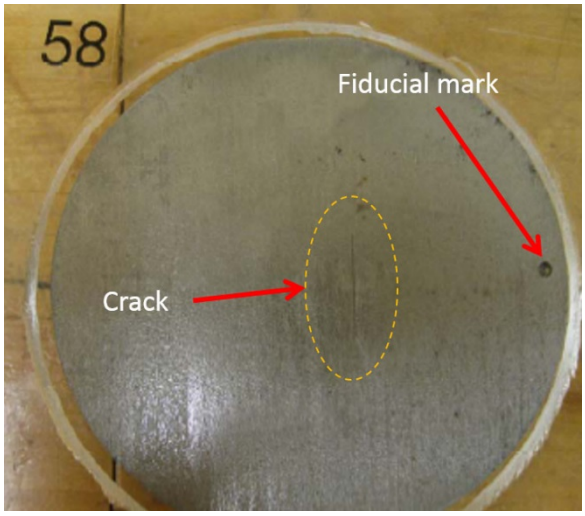


Figure 4-1 Example of Specimen Used for Phase I Mini-RRT

The parametric study specimens, in contrast, were rectangular; their primary purpose was to test the ability to fabricate and use cracks that simulated tight, multi-faceted flaws such as thermal

fatigue cracks or SCC. A secondary purpose was to test the effect of length on detectability using typical RVT instrumentation. Given these two objectives, both stainless steel and ceramic specimens were developed for the parametric study.

The different finishes of these specimens (matte finish and as-received) tested the effect of surface finish on detection capability. Stainless steel specimens with as-received finish had surface conditions typical of rolled steel plates (shiny, limited texture, with the potential for several surface scratches). Ceramic as-received specimen surfaces were representative of those typically found in BWR internal components (reddish color simulating some types of oxide deposits and somewhat rough texture). Matte-finish on selected stainless steel specimens was obtained by media-blasting the specimens using stainless-steel balls. Overall, the specimens were fabricated to mimic conditions prevalent in BWR and PWR reactor internals.

The specimens used in the mini-RRT included several surface scratches near cracks; in some instances, the scratches went through the cracks. Several specimens (with and without surface scratches) were included with no cracks to help determine the FCR.

Two fundamental parameters (crack length and COD) were used in the specimen and RRT design process. For the mini-RRT, the distribution of COD and lengths is shown in Figure 4-2. Table 4-2 shows a summary of the COD and length ranges for flaws used in the parametric study. Section 4.1.2 describes the COD quantification procedure in greater detail.

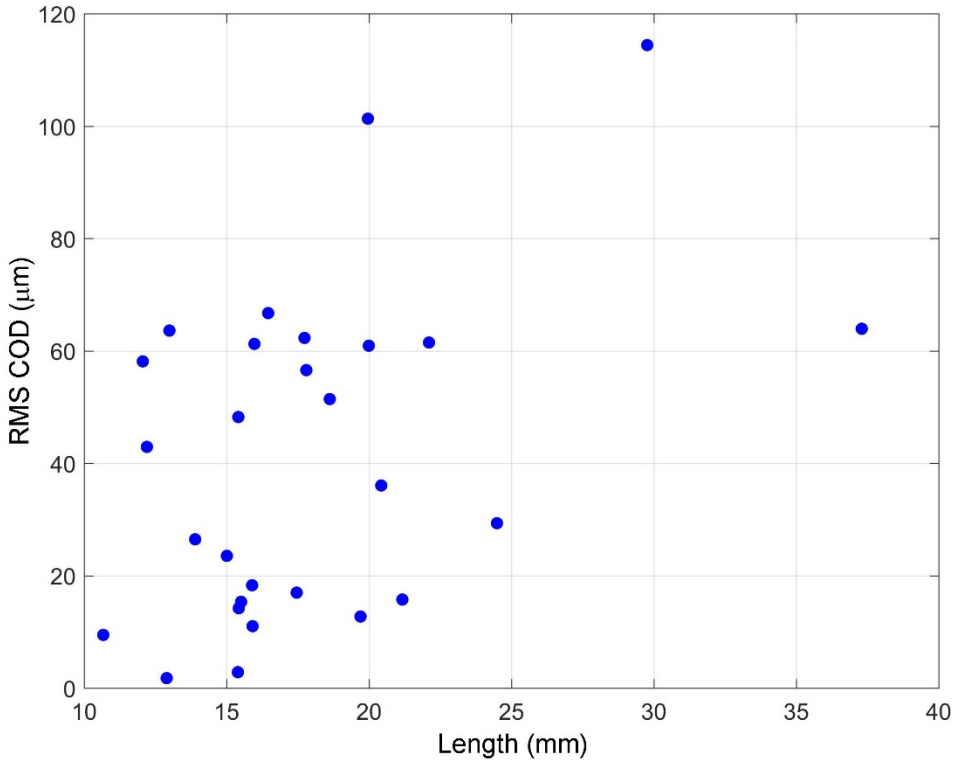


Figure 4-2 Distribution of Crack Lengths and COD for Phase I Mini-RRT Specimens

Table 4-2 Summary of Length and COD Ranges for Phase I Parametric Study

Material	Crack Type	Range of Lengths, mm (in.)	Range of COD, μm*	Surface Condition
Stainless Steel	Mechanical fatigue	~9–30 (~0.35–1.2)	~9–19	Matte
Stainless Steel	Laser notches	~5–35 (~0.2–1.38)	~22	Matte
Ceramic	Laser notches	~5–35 (~0.2–1.38)	~14, ~18	As received

*To convert microns to inches, multiply microns by 0.00004.

SCC in reactor internals (Luk 1993; Ware et al. 1999) appears to be one of the major aging-related degradation mechanisms, and was the focus of the specimen design effort with the goal of producing cracks that mimicked morphologies encountered during inservice examinations. The target size distributions for cracks relied on available references documenting the range of these parameters for SCC from field studies (Wåle 2006), and attempted to mimic the field distributions of these parameters to obtain as realistic a test as possible. Previous studies (Cumblidge et al. 2004; Cumblidge et al. 2007) on the capabilities and limitations of RVT instrumentation were also leveraged where necessary to determine the upper and lower limits for these distributions.

PNNL fabricated a majority of flaws in the stainless steel specimens using a method that employed a starter notch on the opposite side of the specimen (the side not visible during tests). The specimens were then mechanically fatigued (via tension-tension loading) and monitored to detect the onset of a fatigue crack breaking through the surface. The load was then controlled and crack length monitored until it reached the approximate target length. The specimens were then stressed (via tension or compression) perpendicular to crack orientation to achieve desired COD values. COD values before this post-stressing process were also recorded for documentation purposes only and not used in the final analysis. A comparison of COD values on a limited number of cracks after post-stressing and after the Phase 1 test did not show any significant changes. As a result, it was assumed that any stresses introduced during specimen handling after fabrication did not contribute to any significant changes in COD. In any case, the analysis method (described in Section 4.2) assigned cracks into defined COD ranges, and any small variations would not affect the final results.

EPRI produced a subset of flaws in the stainless steel specimens using a pulsed laser. This method used preset crack morphologies as input to the laser cutting process and resulted in shallow notches that mimicked these morphologies. The advantage of this approach was the ability to repeatedly generate simulated cracks that represented arbitrary morphologies; the same approach was also used to generate the flaws in the ceramic specimens. The disadvantage of this approach was the relatively consistent COD from one end of the flaw to the other; this notch-like appearance showed high contrast against the ceramic specimen background. At the same time, the laser spot size and resultant specimen heating produced discoloration along the notch edge which would not be found in nature and could potentially bias the detectability of the flaw. This limited the minimum COD that could be achieved in both ceramic and stainless steel specimens using the laser method.

4.1.2 Quantification of COD

As part of the RVT Phase I tests, measurements of COD were performed using optical microscopy images at both EPRI and PNNL. These measurements, following generally accepted procedures at both institutions, were typically recorded at a small number of locations along the crack.

During measurements, it was found that the COD of the fatigue cracks in stainless steel specimens generally did not vary dramatically over 1–2 mm (0.04–0.08 in.) of length, especially around the center of the crack. Exceptions appear to be on the tighter cracks (maximum COD less than 10 microns [0.0004 in.]). This lack of variability was likely due to the fabrication procedure, and indicated a need to utilize a different fabrication procedure for subsequent phases, if the intent was to simulate SCC.

A mechanism was necessary to convert this set of COD measurements into a single number to be used in subsequent analyses. For this purpose, the root mean square (RMS), maximum, median, and mean COD values were compared. These metrics generally showed good correlation where median, mean, and RMS values appear to be in the same “ballpark” or COD bin. The maximum COD value did not capture variability along the length of the crack and was seen to be susceptible to conditions such as grain dropout that result in a short (1–2 mm [0.04–0.08 in.]) segment with a large COD. As a result, the RMS COD was selected for subsequent analyses.

The metrics attempt to capture all the information about the crack (including variability in COD across the length of the crack). However, this depends on which locations are included in the calculations and which are not. For instance, including COD measurements from the ends of the crack tends to reduce the mean, median, and RMS values, while more measurements in the center of the crack would increase these values.

In most instances, the locations of these measurements along the crack length were difficult to quantify. The cause for concern based on these data was that the COD measurements may not be repeatable, especially if the COD varied along the crack length. This led to development of a consistent approach for quantifying COD of complex cracks that is described in greater detail in Appendix C.

The modified procedure described in Appendix C used a significantly higher density of measurements taken at locations referenced to the crack tips, which ensured repeatability of the COD measurements. This approach was used for computing the COD in Phases II and III.

4.1.3 Test Methodology

Initially, the test methodology consisted of a mini-RRT followed by a parametric study. The mini-RRT consisted of several ISI service providers and camera vendors participating in a “blind” test where the true condition of the specimens was not revealed. The participants were asked to determine if a specimen contained a crack and the approximate location, orientation, and length of the crack. All participants were allowed to retake the test multiple times although time limitations resulted in some participants being unable to retake the test. The order in which the specimens were presented changed for each test, but the participants were not informed of their performance results after each test. In each test, the specimens were also assigned a random identification label that did not reveal the true identity of the specimen. The identification label was changed for each test to ensure that test takers could not identify the specimen.

Each test was performed underwater, with ambient lighting limited by using a black-out tent. Test subjects were asked to limit their lighting options to those available on the participant-supplied cameras. Cameras were mounted on a motor-controlled scanning bridge and the specimens were placed underwater. Each participant was asked to scan across the specimen set and, after inspecting a specimen, record their findings on a provided data sheet.

Prior to each test, participants were asked to perform a resolution check using a standard of their choice. The standard used in each case turned out to be the ASME character set standard that is commonly used for a VT-1 resolution check. This standard uses lowercase characters without an ascender or descender (e.g., a, c, e, o) and the maximum height for resolution demonstration is 1.1 mm (0.044 in.) (ASME 2015c).

In contrast, the parametric study was not a blind test; the test subject (staff at EPRI) was aware of the true state of the specimens. As in the mini-RRT, the parametric study was performed underwater using a motor-controlled scanning bridge and ambient lighting restricted by using a black-out tent. Several cameras were used, including a HD camera. However, other equipment (monitor type, monitor resolution and color setting, and room lighting) used the same configuration for all parametric tests. Participants in the parametric study used a qualitative scoring scale to determine if a crack was detectable using the equipment. The assigned scores were 0 (flaw not detected; no indication of flaw in data record), 1 (poor detection; detection not likely without prior knowledge of location), 2 (moderate detection), or 3 (good detection; easy to detect).

4.2 Analysis Methodology

The analysis of the mini-RRT data was performed manually. Initially, grading units were defined on each mini-RRT specimen as consisting of a quarter of the specimen. However, using these grading units was difficult due to inconsistent camera orientations, inconsistent reporting of fiducial marker location, and inaccurate crack length and location reporting. Further, many cracks tended to overlap multiple grading units, requiring accurate reporting of crack location by the participant for accurate grading. As a result, in Phase I the decision was made to grade based on whether a flaw was detected or not.

The grading, where necessary, was augmented using recorded video provided by the participants and verified by an independent grading assessor. Data, collated by specimen identifications used in each test and by participant, were used to determine the detection and FCRs for each test. The detection and FCR were separated by participant, as a function of COD and length. COD was broadly categorized by binning cracks into one of four categories (0–20 μm, 20–40 μm, 40–100 μm, and greater than 100 μm [0–0.0008 in., 0.0008–0.0016 in., 0.0016–0.004 in., and greater than 0.004 in.]) for the purposes of computing detection rates. The distribution of cracks in these bins was not uniform, with the most cracks in the 0–20 μm (0–0.0008 in.) and 40–100 μm (0.0016–0.004 in.) bins. Only one crack had a COD larger than 100 μm (0.004 in.).

The number of false calls on each specimen and the total number in each test were reported. False calls included any calls on blank or flawed specimens that were not in the same location as the flaw.

The results for detections and FCRs were further examined to determine if factors such as camera type, use of HD cameras, and other parameters affected the results.

Parametric study data were analyzed manually, using recorded data only (analysis using live data was performed at EPRI), and according to a subjective scale of 0–3, with 0 corresponding to no detection and 3 corresponding to easy detection.

4.3 Phase I Results

4.3.1 General Findings from Mini-Round Robin

Detection performance and FCRs from Phase I (including the parametric study) were analyzed as a function of a number of parameters, including COD, length, camera type, field of view of the camera, and operator training level.

Figure 4-3 summarizes the overall detection rate from the mini-RRT, averaged over all trials and all participants, as a function of the COD. As seen from this graphic, the overall detection rate increases as the COD increases and there is no significant difference in detection rate between the various computation methods (RMS, median, maximum). Note the detection rates when using RMS and median to compute COD are the same.

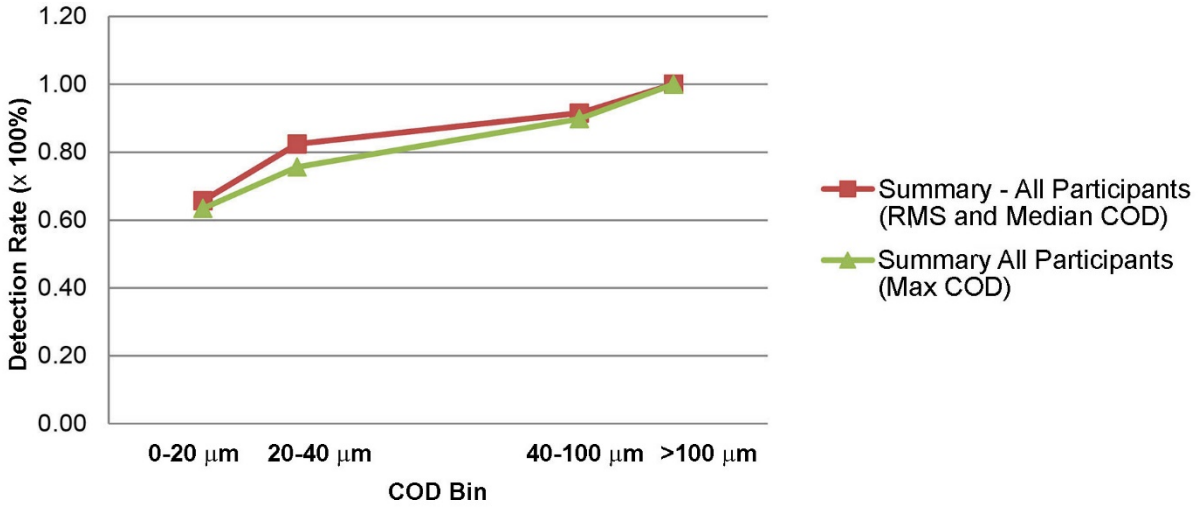


Figure 4-3 Detection Rate (total, over all trials) as a Function of COD. Detection rates as a function of RMS and median COD are identical.

Figure 4-4 shows the detection rate for each trial, where a trial consisted of a single test administered as part of the mini-RRT, as a function of the same COD ranges. This breakdown by trial shows interesting variations, particularly for the smaller COD cracks. An evaluation of the underlying data indicates that potential factors influencing this result include participant experience as well as a potential learning effect that may have occurred as a result of participants being allowed to take the test multiple times.

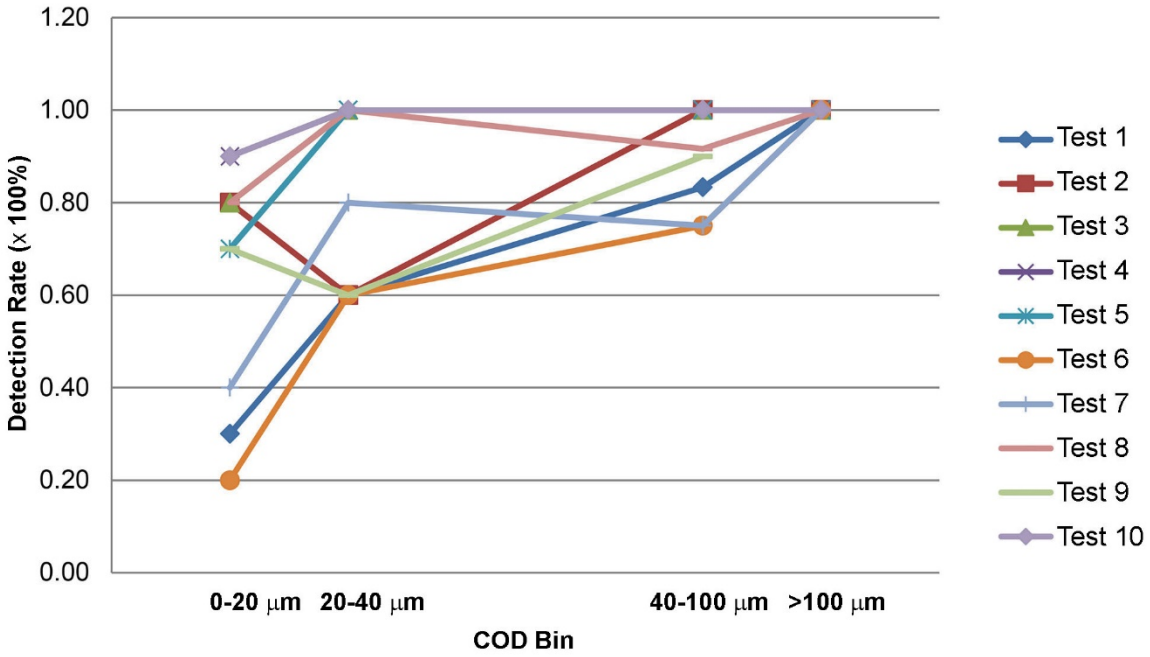


Figure 4-4 Detection Rate as a Function of RMS COD for Each of the Ten Tests (Trials). Data from some of the tests are identical and the corresponding data plots are hidden by the data from other tests.

Figure 4-5 shows clear differences in the average detection rate between experienced participants vs. other participants. The detection rate, when accounting for repeated trials, also shows an improvement. Anecdotally, during debriefing, participants indicated that their familiarity with the specimens increased as they were given multiple opportunities to go through the tests, and that they were better able to adjust to the surface condition and lighting needs to detect cracks.

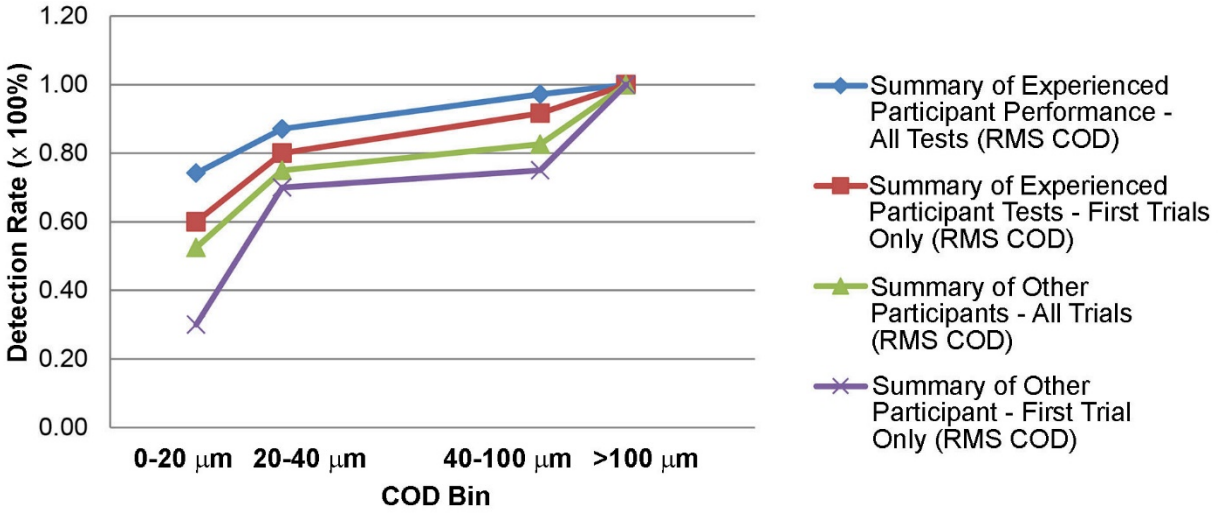


Figure 4-5 Detection Rate (organized by participant experience) as a Function of RMS COD. Because some participants took the test multiple times (although using different cameras), the results are further separated by combining only the first trials for each participant vs. all trials for the participants.

Analysis indicated the overall FCR was low when measured in false calls per specimen. However, there was again a quantifiable difference in the FCR based on the participant's experience, with the experienced participants who perform RVT examinations routinely having fewer false calls overall. This is illustrated in Figure 4-6, along with the detection rates. A possible cause of this result is that experienced participants are trained to assess an indication possibly using multiple viewing angles, and adjusting the camera zoom and lighting, enabling them to more effectively discriminate between flaws and other non-reportable indications such as surface features. In practice, this may mean a lower FCR in field examinations.

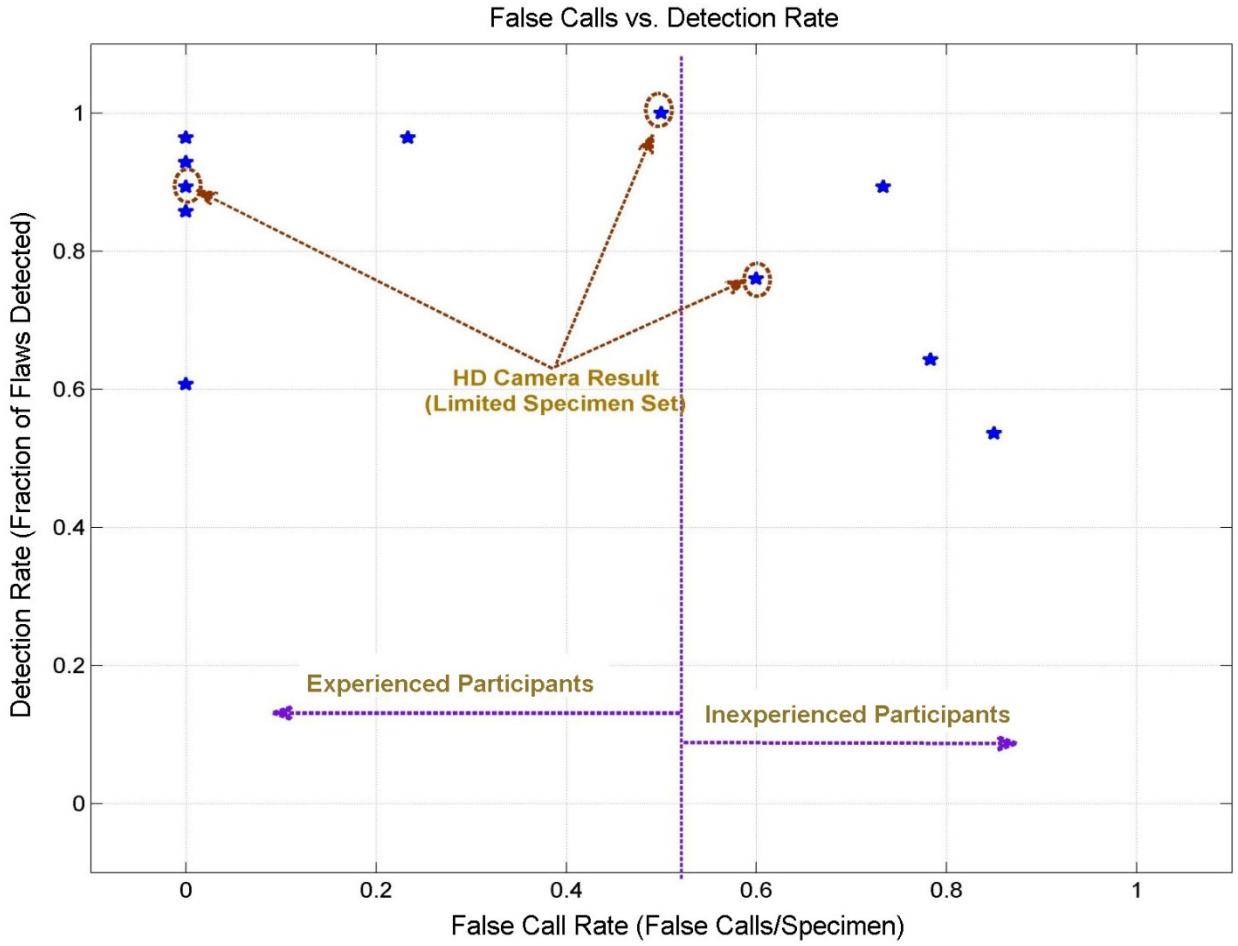


Figure 4-6 Detection vs. FCRs for Each Test

Note that as-received surface condition of the specimens did not appear to significantly impact the test results, as the overall detection rates were fairly high, and FCRs were generally manageable at less than one per specimen.

Examination of the video data showed that, in many cases, smaller COD cracks (which were not identified by the participant) are visible, indicating the cameras, even standard definition ones, are capable of viewing many of the small cracks. However, these were often incorrectly dismissed as surface imperfections. Camera pan-tilt-zoom capabilities were important to detection, but tube-type cameras with no pan-tilt-zoom appeared to be equally capable in the hands of an experienced operator. Lighting, when applied properly, made a significant difference in the visibility of a flaw as well as in discriminating between a flaw and non-flaw.

An interesting mini-study in Phase I included an evaluation of the use of HD cameras. While these were not used in the field at the time of the test, HD cameras were applied in a limited manner in Phase I and appeared to better resolve small flaws. However, FCRs may increase with the use of HD cameras. Data from Phase I were insufficient to determine this conclusively, as only two teams used HD cameras on a subset of the specimens.

The first HD team, with an experienced RVT analyst, performed well with regard to flaw detection; on the other hand, the FCR was high when using standard definition cameras by the same team and analyst. The improvement in detection capability could also be due to the experience gained on the specimens as the HD camera was used on the last round of testing by this team. The other HD team, which did not have an experienced RVT analyst, had high numbers of false calls in addition to high detection. Again, experience with the specimens and flaw types may be a factor in increased detection.

Interestingly, in both cases the detection rate for cracks with COD in the 20–40 micron (0.0008–0.0016 in.) range fell slightly when using an HD camera. The limited amount of data was insufficient to determine if this result was statistically significant.

4.3.2 General Findings from Parametric Study

Analysis of the parametric study data pointed to several interesting findings. First, there was an apparent difference in detection capability between the live and recorded data sets, which was attributed to an apparent difference in clarity between the two data sets. This difference in quality was likely due to use of file compression in the recorded data, which is a common practice when recording and saving video. The quality of recorded data was apparently dependent on both the camera system and specimen texture, where less texture resulted in greater compression and poorer video quality for the purposes of flaw detection.

Beyond this general observation on the potential for reduced data quality in recorded video, the ability to detect cracks was apparently dependent on the flaw length, shape/tortuosity, COD variation along the length of the crack, local grain structure variations, and variation in surface texture. For a given length, crack detection was observed to be more challenging as the COD decreased. Figure 4-7 shows an example of this, for cracks (represented by diamonds in the figure) and notches (circles in the figure) in stainless steel specimens. Similar performance was observed in ceramic specimens, although the overall detection rates were lower. It was not clear if this reduction was due to inherent limitations in RVT in detecting flaws in specimens with little to no texture, as was the case in the ceramic specimens, if this was a result of the reduced data quality observed in the recorded data, or both.

Figure 4-7 shows the detection score averaged over all tests from a single camera, which varied the camera distance from the specimen and the field of view. There does not appear to be a large variation between the data from the different cameras, though longer flaws are easier to detect. However, the data appear to indicate that the notches may be easier to detect than cracks. Possible reasons for this include the somewhat uniform nature of the COD along the notch as well as the larger lower COD limit on laser-fabricated notches in stainless steels.

The results overall seemed to indicate correlation of detection performance with both COD and length, with little to no variation among cameras.

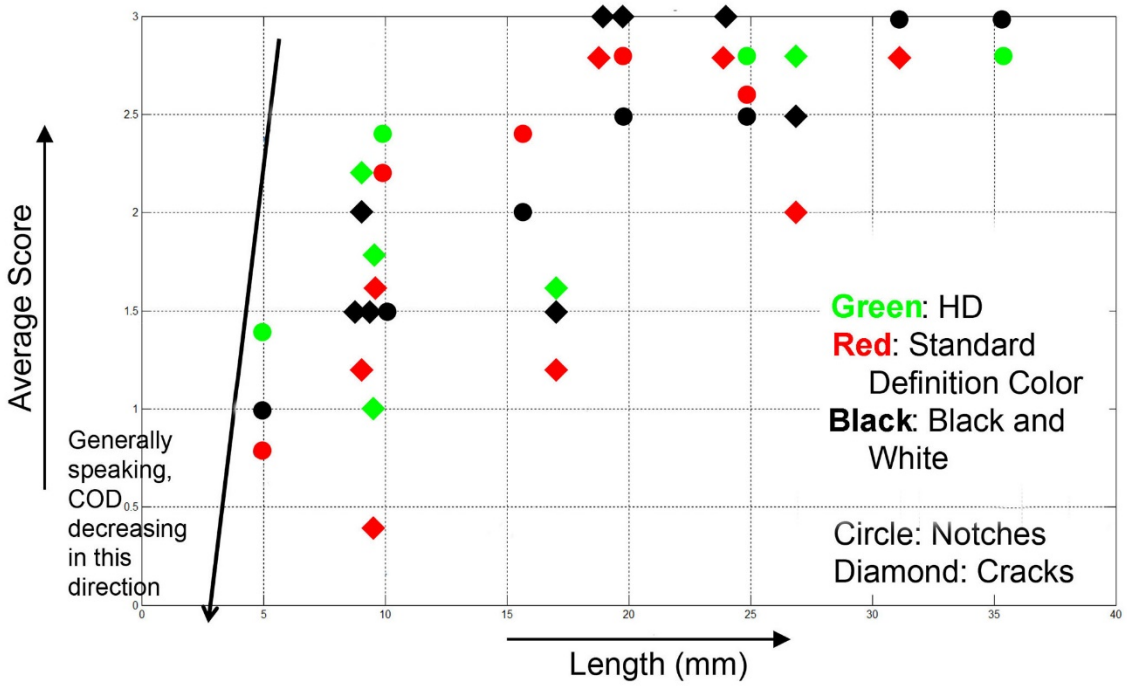


Figure 4-7 Average Detection Score vs. Length for Cracks and Notches in Stainless Steel

4.3.3 Outcomes from Phase I

The results of Phase I pointed to the need for improvements in specimen fabrication and the test protocol, and a potential change in the distribution of flaw lengths and COD for Phase II. Phase I indicated that both stainless steel and ceramic specimens may be appropriate for use in Phase II, with cracks fabricated in the stainless steel specimens designed to overcome the limitations of the laser-fabricated notches in stainless steels. The Phase II protocol needed to carefully consider the impact of using recorded data for analysis and require a mechanism to capture details of the instruments and instrument settings used by the participants.

Phase II was designed to factor in several of the key findings from Phase I. These are summarized below and organized according to whether the key finding was attributed to specimen design and fabrication, flaw characteristics, equipment, protocol, or personnel.

4.3.3.1 Specimens

- Phase II used flat plates with weld crowns to simulate small sections of welds and heat-affected zones in typical internal components.
- Blank specimens were given unique labels, similar to the flawed specimens in both Phases II and III. This was conducted to aid in identifying and monitoring any correlations between false calls by multiple teams on the same specimen.
- Several Phase II specimens contained multiple flaws, to ensure that operators were not conditioned towards a one-sample-one-flaw testing protocol.

4.3.3.2 Flaws

- Phase II included a larger number of cracks with smaller COD that had a median of 10–20 μm COD distribution (Wåle 2006). A small number of larger flaws were included to ensure an appropriate statistical flaw distribution.
- Phase II included a range of flaw length-COD combinations to better quantify the effect of flaw length.
- Flaw mechanism may also play a role in detection performance; however, Phase II only simulated the crack shapes of mechanisms such as SCC and thermal fatigue cracks using mechanical fatigue cracks and laser-cut notches.

4.3.3.3 Phase II Protocol

- To avoid repeating the same specimen protocol and the scenario where the analyst may become familiar with individual specimens, Phases II (and III) restricted each team to one test with all the specimens.
- Grading units were better defined for Phase II.

4.3.3.4 Phase II Participants

- Phase II focused on inspection teams consisting of personnel with field experience in IVVI and included Level III analysts only.

4.3.3.5 Findings Not Addressed in Phases II and III

- While there was a recommendation for Phase II to include more realistic conditions, such as specimen surfaces representative of actual reactor pressure vessel (RPV) internals and camera rigging similar to that used in industry, constraints on resources and time and access to a suitable RPV test facility precluded this in-field realism.
- The limited data from Phase I did not identify a clear advantage to using HD cameras. Both HD and standard defiinition cameras appeared to be capable of imaging most of the smallest cracks in Phase I. The test did not provide sufficient data to determine if HD cameras would result in an increase in false calls. Given the focus on using field-deployed instrumentation in Phases II and III, HD cameras were not tested again for the balance of the assessment. This is an issue that remains open and may need to be re-evaluated, perhaps using a parametric study.
- The effect of training could not be evaluated further given resource constraints. Such a study could require a protocol that includes several analysts using the same camera and evaluating specimens multiple times in different order. One or two analysts with multiple cameras on the same specimens, to assess the effect of camera/lighting variation, may also be necessary. This type of test limits the number of changing variables so the effect of any one factor can be isolated and clearly identified.
- The experience level of the analyst appeared to play a large role in the VT results. A program to train and provide experience may be necessary. As discussed in Section 7, this recommendation was reinforced by the findings of Phase III.

5 PHASE II OVERVIEW AND RESULTS

5.1 Overview

The results of Phase I pointed to the need for a careful design of specimens for Phase II and included the following key recommendations:

- Include a representative population of cracks where a large fraction of these cracks consist of COD around 20–40 μm.
- Limit the number of times a participant can take the test and restrict participants to ISI service providers.

Several other recommendations, such as a larger base of experience levels for inspectors and different types of cracking, had to be eliminated for resource reasons. The test design for Phase II was focused on quantifying the test performance in terms of capabilities and limitations with respect to “realistic” cracking and as a function of parameters that are independent of the type of crack. As a result, it is expected that the results can be applied to determine a baseline performance for RVT.

5.2 Experimental Design

Phase II formally evaluated the capabilities of visual testing methods to detect surface-connected cracks in BWR and PWR reactor components. Test specimens that represent only BWR components were manufactured from ceramic material, while those representing both BWR and PWR components were manufactured from stainless steel. The test specimens consisted of plates approximately 250 mm (9.8 in.) in length/width, with a weld (or simulated weld crown) at the midpoint. Cracks fabricated in these specimens were oriented either parallel to the weld direction (referred to as “circumferential” or “circ” hereafter) or transverse to the weld direction (referred to as “axial”).

5.2.1 Specimens

This section presents the statistical design for the RV-RRT specimens. Two designs were initially considered, requiring either 30 or 45 test specimens to be manufactured for the round robin. The original 45 test specimen design was modified slightly to use 44 test specimens, and is described here.

The objective was to construct test specimens and cracks that mimic material inspected in the field. PNNL and EPRI identified a number of parameters as important descriptors of in-field conditions (Table 5-1). For each parameter, an appropriate frequency of occurrence in the field (in-field frequency) was assumed.

Table 5-1 Design Parameters for Phase II Test Specimen Material

Reactor Type	Weld Crown	Scratch Orientation	Scratch Density	Surface Roughness	In-field Frequency
Ceramic	Ground	45-degrees	0.25/mm (0.01/in.)	Typical	0.100
Ceramic	Ground	Parallel	0.25/mm (0.01/in.)	Typical	0.025
Ceramic	Unground	Parallel	0.25/mm (0.01/in.)	Typical	0.124
Ceramic	Unground	Parallel	None	Typical	0.001
Stainless steel	Ground	45-degrees	0.25/mm (0.01/in.)	Typical	0.480
Stainless steel	Ground	Parallel	0.25/mm (0.01/in.)	Typical	0.120
Stainless steel	Unground	Parallel	0.25/mm (0.01/in.)	Typical	0.120
Stainless steel	Unground	Parallel	None	Typical	0.030

The population of test specimens was designed to reproduce these frequencies as closely as possible. Other factors were also considered during the design of the test specimens. These included:

- **# Flaws Per Specimen**: Average of 2, maximum of 5.
- **Distribution of Cracks over Test Specimens**: 15% with no cracks, 27% with 1 crack, 27% with 2 cracks, 18% with 3 cracks, 9% with 4 cracks, 5% with 5 cracks.
- **# Blank Grading Units**: Number of blank grading units (GUs) should be at least 25% of number of cracked GUs, with 10 blank GUs located in blank test specimens.
- **Nominal Grading Unit Size**: Crack dimension plus 1 cm (0.39 in.) in all directions.
- **Grading Unit/Crack Placement**: Randomly located within test specimen but with minimum distance of 4 cm between grading units.

The Phase II test design included cracks in 15 ceramic specimens and 30 stainless steel specimens, for a total of 45 specimens. However, only 44 specimens were fabricated—15 ceramic and 29 stainless—due to fabrication difficulties on one stainless steel specimen that was designed to contain a single crack with a small COD. Tables 5-2 and 5-3 summarize the as-built crack matrix for all the stainless steel specimens and ceramic specimens used in Phase II, respectively. These tables describe the distribution of the length and COD combinations for the different weld conditions and orientations of the cracks. The total number and distribution of cracks in the 44 specimens was deemed to be statistically sufficient to meet the objectives of Phase II.

Table 5-2 Flaw Matrix Table for Flaws Contained in 29 Stainless Steel Specimens Used in Phase II

Material	Weld Crown	Flaw				# Flaws (built)
		Fabrication Type	Orientation	Length	Width	
Stainless Steel	As welded	Fatigue	Circ	Small	Small	4
				Small	Medium	2
				Small	Large	2
				Medium	Small	3
				Medium	Medium	2
				Medium	Large	4
				Large	Small	1
				Large	Medium	2
				Large	Large	7
			Axial	Small	Small	0
				Small	Medium	0
				Large	Small	1
				Large	Medium	0
				Medium	Small	5
				Medium	Medium	2
Total (stainless steel, as welded)						**35**
Stainless Steel	Ground	Fatigue	Circ	Small	Medium	5
				Medium	Small	4
				Medium	Medium	3
				Medium	Large	0
				Large	Medium	0
			Axial	Small	Small	2
				Small	Medium	1
				Large	Medium	1
				Medium	Small	4
				Medium	Medium	0
Total (stainless steel, ground)						**20**

COD: small = 5 to 20 µm (0.2 to 0.8 thou.), medium = 21 to 40 µm (0.82 to 0.1.57 thou.), and large > 40 µm (0.1.57 thou.). Length: small ≤ 15 mm (0.6 in.), medium > 15 mm (0.6 in.) and < 26 mm (1.02 in.), and large ≥ 26 mm (1.02 in.).

Table 5-3 Flaw Matrix Table for Flaws Contained in 15 Ceramic Specimens Used in Phase II

		Flaw				
Material	Weld Crown	Fabrication Type	Orientation	Length	Width	# Flaws (built)
Ceramic	Unground	Laser	Circ	Small	Medium	7
				Medium	Medium	5
				Large	Medium	2
				Small	Small	5
				Medium	Small	4
				Medium	Large	1
			Axial	Medium	Small	1
				Small	Small	1
				Large	Small	0
				Medium	Large	1
				Small	Large	0
				Large	Medium	0
				Medium	Medium	3
Total (ceramic, unground)						**30**

COD: small = 5 to 20 µm (0.2 to 0.8 thou.), medium = 21 to 40 µm (0.82 to 0.1.57 thou.), and large > 40 µm (0.1.57 thou.). Length: small ≤ 15 mm (0.6 in.), medium > 15 mm (0.6 in.) and < 26 mm (1.02 in.), and large ≥ 26 mm (1.02 in.).

Table 5-4 summarizes the 44 test specimens produced for the RRT in terms of the number of areas with surface features (scratches, grinding, or scuff marks) in test specimens and weld crown condition. Tables 5-5 and 5-6 further break down the distribution of cracks based on orientation of cracks and surface features (Table 5-5), and the location of cracks (Table 5-6). The different types of cracks are more uniformly distributed throughout the stainless steel material than the ceramic (Table 5-6). Further, in stainless steel, no axial cracks are present in the heat-affected zone (HAZ).

Table 5-4 Summary of Test Specimens

Material	Weld Crown	Number of Areas with Surface Features	Number of Specimens
Ceramic	NG	0	11
Ceramic	NG	1	3
Ceramic	NG	2	1
Stainless steel	G	1	1
Stainless steel	G	2	2
Stainless steel	G	3	2
Stainless steel	G	4	2
Stainless steel	G	5	1
Stainless steel	G	8	1
Stainless steel	NG	0	2
Stainless steel	NG	1	4
Stainless steel	NG	2	6
Stainless steel	NG	3	5
Stainless steel	NG	4	2
Stainless steel	NG	5	1
Total			**44**

G = ground; NG = not ground (i.e., as welded)

Table 5-5 Summary of Cracks and Surface Features in the Test Specimens

Flaw Type	Material	Crown	Orientation	Number of Flaws
Crack	Ceramic	NG	A	6
Crack	Ceramic	NG	C	24
Crack	Stainless steel	G	A	8
Crack	Stainless steel	G	C	12
Crack	Stainless steel	NG	A	8
Crack	Stainless steel	NG	C	27
Total Cracks				**85**
SF	Ceramic	NG	A	1
SF	Ceramic	NG	C	4
SF	Stainless steel	G	A	10
SF	Stainless steel	G	C	22
SF	Stainless steel	NG	A	19
SF	Stainless steel	NG	C	25
Total SF				**81**

A = axial; C = circumferential
G = ground; NG = not ground; SF = surface feature (scratched area)

Table 5-6 Flaw Count in VT Round Robin by Location, Flaw Type, and Material

	Ceramic		Stainless Steel	
Flaw Location	A	C	A	C
GrWeld	0	0	6	0
HAZ	4	9	0	23
InSF	2	15	4	15
NGWeld	0	0	6	1

GrWeld = ground weld crown
HAZ = heat-affected zone (includes base metal)
InSF = crack in surface feature
NGWeld = not ground weld crown (i.e., weld crown in as-welded condition)
A = Axial
C = Circumferential

5.2.2 Crack Size Distributions

Figure 5-1 shows the distribution of crack sizes (COD vs. length) for the various types of cracks in ceramic and stainless steel specimens. Note that the axis labels for all four sub-plots in Figure 5-1 are the same, and the labels from two of the four plots have been dropped to avoid cluttering the plots. The values for data points in these plots may be determined by examining the axis labels in adjacent plots. The distribution indicates a correlation between COD and length (chiefly associated with the extreme values). However, determining which of these two factors influences the POD should still be possible if the effect is strong.

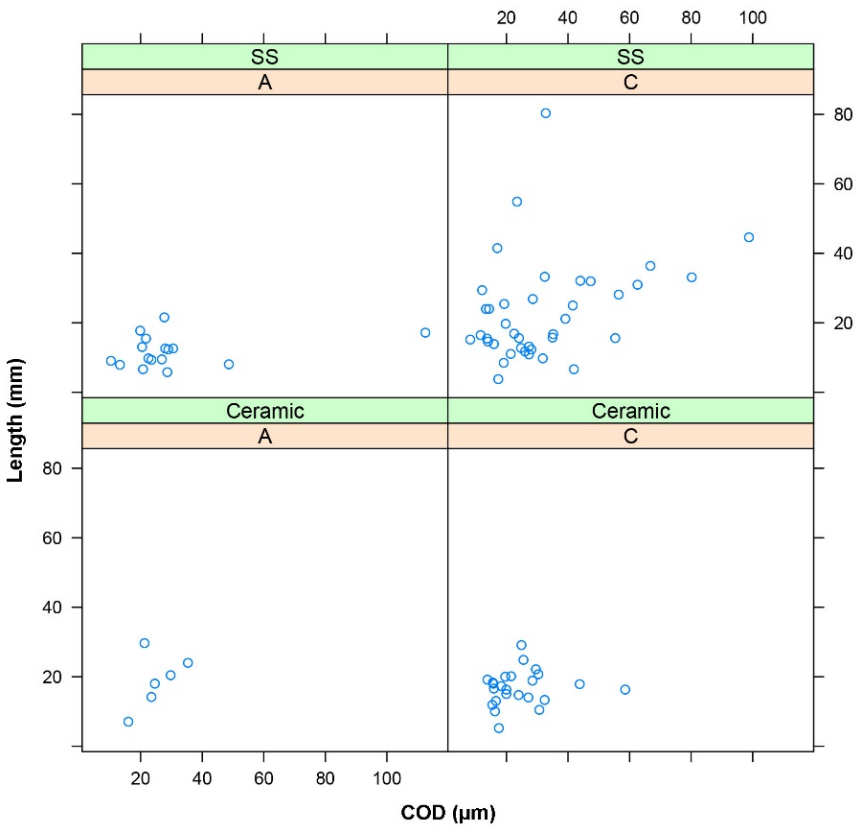


Figure 5-1 Flaw Dimensions Organized by Orientation (Axial, Circumferential) and Specimen Type (Ceramic, Stainless Steel)

5.2.3 Test Methodology

The test methodology for Phase II used a round-robin approach. Five teams from IVVI service providers participated in a blind test where the true condition of the specimens was not revealed. The participating teams were asked to determine if a specimen contained a crack, and its approximate location, orientation, and length, although their ability to accurately estimate the length of the crack was not evaluated. Rather, the length estimates were used only to determine the extent of the crack for automated grading of crack detection.

Unlike Phase I, each team was only allowed to take the test once, using all specimens. Analysis was restricted to using live data only, although all data were recorded for the test administrators (PNNL and EPRI) to use to assist in answering any questions regarding detection during the subsequent grading process. In each test, the specimens were assigned a random identification label that did not reveal the true identity of the specimen. Each of the participating teams was also assigned a random identification code (ALYJ, BMXR, CIWN, DOYP, and EQZH) to maintain confidentiality and anonymity of the participating teams.

As with Phase I, each test was performed underwater, with ambient lighting that was limited using a black-out tent. Participants were asked to limit their lighting options to those available on the participant-supplied cameras. Specimens were placed underwater. The cameras were mounted on a motor-controlled scanning bridge and the inspection was performed underwater. Each team inspected one specimen and recorded their findings on the provided data sheet, after which the

specimen was removed from the tank and replaced with a different specimen. This process was repeated until all specimens were inspected. The start and end times for the examination of a specimen, including the time to document any indications, were recorded on the data sheet.

Prior to each test, the teams were asked to perform a resolution check using the ASME character standard (ASME 2015c). Each team was also asked to provide a copy of their inspection procedures (which were provided to PNNL under non-disclosure agreements), and to perform any other necessary pre-inspection steps as required by their procedures prior to taking the test.

The complete test protocol for Phase II is presented in Appendix A.

5.2.4 Excluded Variables

The test specimens used in Phase II represent welds similar to that found in some reactor internal components. The specimens represent two specific colors (patina), natural stainless steel and reddish tints, but may not necessarily be representative of the diverse color variations of internal RPV surfaces. The effect of other configurations and surface conditions on test results was beyond the scope of this RRT. Finally, the RV-RRT was not designed to assess certain variables that may impact detection performance in remote visual testing. Some of these are:

- Lighting options
- Oxide build-up on internal components
- Thermal distortion
- Water currents and clarity
- Radiation effects on camera video quality
- Limited accessibility (component configurations and camera size)
- Monitors and camera systems
- Camera delivery systems
- Personnel qualification levels
- The angle of view limits for ASME Code VT-1 examinations
- Video compression algorithms

5.3 Data Description and Grading

As described earlier, the RRT included 44 test specimens containing 85 cracks and 81 areas with surface features. All five teams inspected all the specimens, producing a total of 220 inspections. Because each team inspected all of the test specimens, the data are considered balanced and it is much easier to compare the effect of different variables (crack COD, length, orientation, and location) using simple POD tables.

The five participating teams used similar inspection procedures and equipment. All have performed field inspections and have the necessary qualifications for performing inservice inspections. However, no data exist to determine if these five teams can be considered to represent an unbiased sample from the field inspector population, given that there are a range of qualifications for personnel who perform ISI. In the absence of any additional information, their average POD/FCP detection performance should be considered to represent a baseline

performance level from which RVT performance may vary, depending on a number of other factors, including inspection qualifications, conditions in the field, and any restrictions on the inspections.

5.3.1 Grading Procedure

The Phase II data were graded using an automated algorithm that compared the reported inspection results for a test specimen to the true-state information for the same specimen. The grading algorithm assumed a crack was detected if the reported location of a crack from the inspection was close (within the grading tolerance) to the true location of a crack.

The grading tolerance was applied to account for minor errors in reported location. Figure 5-2 displays the relationship between grading tolerance and POD for the data from Phase II. The figure shows that a grading tolerance of 10 mm (0.39 in.) accounts for minor location errors in the data, with the POD in ceramic and stainless steel specimens becoming relatively stable beyond this value.

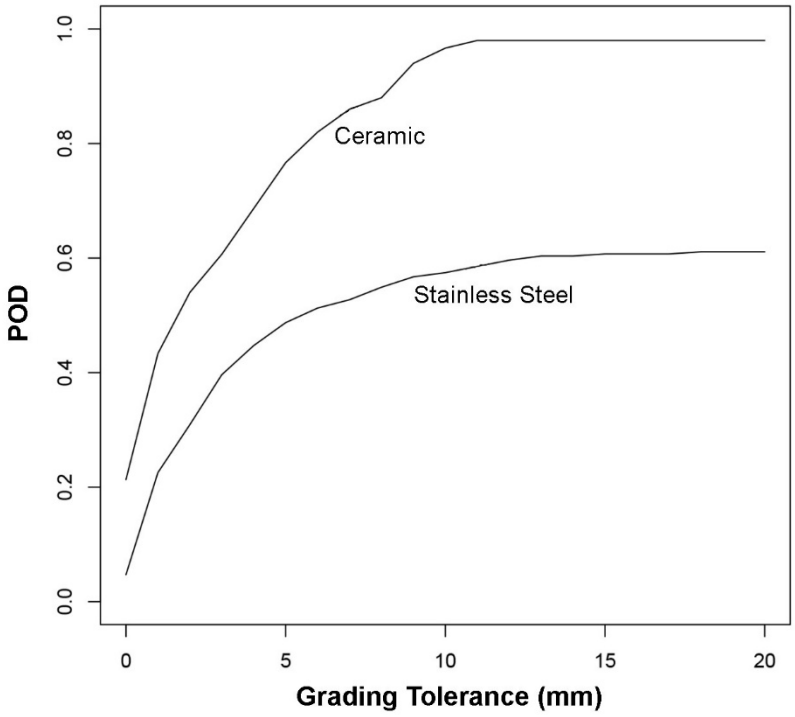


Figure 5-2 Plot of POD vs. Grading Tolerance

5.3.2 Recording Errors

The automated grading algorithm, while convenient, has three potential issues. First, in cases where the true crack is close to a surface feature (within the grading tolerance), it cannot determine if the analyst reported the surface feature as opposed to the crack. Second, in cases where two or more indications are reported by the analyst close to each other (again, within the grading tolerance), the grading process may associate more than one indication with the crack. While many instances of this type of error can be handled through appropriate rules in software, it is possible that one or more exceptions to the rule may exist. Finally, errors in reported location of indications may exist in the data. To address all of these issues, the automated grading was

augmented with a manual review of the analysis results. This manual review resulted in the identification and correction of several gross recording errors made by the participants. These corrections were included in the analysis of grading tolerance shown in Figure 5-2.

The manual review indicated that in 28 cases, the participating teams had recorded incorrect coordinates for the reported indications. While the corrected coordinates were used in subsequent analyses, the errors were used to estimate a probability for gross recording errors (Table 5-7).

Table 5-7 Gross Recording Error Probability in VT Round-Robin Inspections

	Specimen Type	
Team	Ceramic	Stainless Steel
ALYJ	0.000	0.023
BMXR	0.033	0.000
CIWN	0.567	0.079
DOYP	0.000	0.057
EQZH	0.080	0.000

While the average error rate in recording indication locations is about 3%, there are variations from this average, with team "CIWN" appearing to exhibit severe coordinate recording problems (e.g., more than 50% of the cracks in the ceramic test specimens had incorrect coordinates). If these results reflect field performance, this implies that reported indication coordinates may be in error some 3% of the time. It also implies that a POD curve that accounts for such errors can never exceed 97%. However, it should also be emphasized that the test was not designed to explicitly quantify location recording errors or errors in quantifying length. These findings on recording errors should therefore be considered as requiring additional controlled tests for quantification.

5.4 Overview of Analysis Methodology

Inspection performance is quantified using POD and FCP. Perhaps the best overview of POD and FCP is given by tables that estimate these quantities for various conditions (orientation, location, etc.). The effect of continuous variables (crack COD and length) is evaluated using logistic regression models. The most basic model used involves crack size:

$$\mathrm{POD}(S) = \mathrm{logistic}(\beta_0 + \beta_1 S), \tag{5-1}$$

where *S* represents crack size (specifically, COD or length). To compare the relative importance of COD and length, both explanatory variables may be included:

$$\mathrm{POD}(\mathrm{COD}, \mathrm{Length}) = \mathrm{logistic}(\beta_0 + \beta_1 \mathrm{COD} + \beta_2 \mathrm{Length}). \tag{5-2}$$

A curve-fitting procedure is used to estimate the unknown parameters (β_0 , β_1, and β_2) in Eqs. (5-1) and (5-2) and estimate the POD curves as a function of the crack size. POD curves were fit with and without false call data. In some cases, inclusion of false call data produced a curve that did not fit all data, indicating that the two-parameter regression model may not be flexible enough to describe POD for the range of crack sizes present in the study.

Appendix D presents an overview of POD modeling and Appendix G describes alternative models for fitting data. The basic finding from this assessment is that, while alternative models such as the Box-Cox model (see Appendix G) may provide a better fit to the data with false calls included, the overall results, such as the crack parameter at which 80% POD is achieved with a 95% confidence level, do not change substantially. As a result, the general findings from Phase II (and Phase III) should be considered representative.

5.5 Phase II Results Summary

Based on the previously described assessments of grading tolerance, a 10 mm (0.39 in.) tolerance was selected and used for the automated grading. As indicated above, this was accompanied by manual review of the data to correct for gross location errors. The data were then analyzed to quantify performance in terms of POD and determine the effect certain important parameters have on POD, such as COD, crack length, and the presence of surface irregularities and surface features.

A summary of these results is provided below, along with a discussion on the potential implications and outcomes leading to Phase III. A detailed summary of the results is included in Appendix E.

5.5.1 General Findings from Phase II

5.5.1.1 Summary of Detection and False Call Rates

Table 5-8 provides an overview of team performance in Phase II. POD in this table is based on all the cracks and does not distinguish performance as a function of the crack COD, length, or location. The POD is, however, separated according to the specimen type (ceramic vs. stainless steel).

Table 5-8 POD, FCP, and FCRs by Team and Specimen Type

	Ceramic				Stainless Steel			
	POD	FCP			POD	FCP		
Team	Crack	FCP (SF)	FCP (Blank)	FCR (FC/M)	Crack	FCP (SF)	FCP (Blank)	FCR (FC/M)
ALYJ	1.00±0.02	0.40±0.22	0.07±0.04	1.12±0.59	0.78±0.06	0.17±0.04	0.05±0.03	0.79±0.45
BMXR	1.00±0.02	0.40±0.22	0.00±0.02	0.00±0.25	0.62±0.07	0.09±0.03	0.02±0.02	0.26±0.31
CIWN	1.00±0.02	0.40±0.22	0.00±0.02	0.00±0.25	0.69±0.06	0.05±0.03	0.05±0.03	0.79±0.45
DOYP	1.00±0.02	0.40±0.22	0.00±0.02	0.00±0.25	0.64±0.06	0.12±0.04	0.06±0.03	1.05±0.50
EQZH	0.83±0.07	0.00±0.12	0.05±0.04	0.75±0.53	0.15±0.05	0.00±0.01	0.02±0.02	0.26±0.41

Crack = cracked grading unit
SF = surface feature; corresponds to FCP in scratched grading units
Blank = blank grading unit
FC/M = false calls per meter

The POD information in Table 5-8 indicates a relatively high capability to detect cracking in materials with surface conditions similar to those in the ceramic specimens used in Phase II.

However, detecting cracking in stainless steel specimens appears to have been challenging. Several possible reasons exist for this difference and are discussed later.

The data from Phase II indicate a significant difference in POD and FCR for team EQZH. It is not clear whether this is because team EQZH is using a more stringent detection threshold than the other teams or for some other reason. The pattern is most noticeable for stainless steel.

In each case (ceramic and stainless steel), two approaches to computing FCP are used. The first approach defined FCP as the probability of calling a crack in a blank (i.e., no cracks are present) grading unit, designated as FCP(Blank) in Table 5-8. The second approach to computing FCP used the probability that a surface feature (i.e., within a scratched grading unit) is called a crack, designated as FCP(SF). Table 5-8 shows that it is much easier to distinguish cracks from blank material as opposed to cracks from surface features.

The FCR is defined as the number of false calls per meter of blank material. The data in Phase II show that the FCR appears to be roughly 1 false call per meter in blank material. It is not clear how this relates to FCRs in field inspections, and whether the test environment provides an incentive for lowering detection thresholds, thereby increasing POD and FCR. Given that data for field FCR are difficult to obtain or quantify, the FCR and POD information should only be considered as a baseline against which subsequent analyses and any information on field performance can be compared to.

The data presented in Table 5-8 includes the standard deviation (or standard error) as a measure of the uncertainty in the POD, FCP, and FCR estimates. The standard deviations are dependent on the sample size (number of cracked and non-crack grading units present in the Phase II test). Increasing the sample size (by increasing the number of cracks, for instance) generally reduces the uncertainty and leads to smaller standard deviations. Two estimates with large overlapping standard deviations may be considered to be statistically similar. In the data presented in Table 5-8, the standard deviations for the different teams are seen to be roughly similar. This is expected as the sample sizes for each team are the same.

5.5.1.2 POD Analyses

The Phase II data were further analyzed to extract the POD as a function of the two crack variables (COD and length). As discussed in Section 5.4 (and in greater detail in Appendix D), extracting the POD as a function of COD or length (or other independent variable) generally uses a regression model to fit the data. In the present case, the logistic regression model was used to compute the POD curves. In all cases, the FCP was used as part of the input to the regression model. While the incorporation of the FCP leads to the POD curve not passing through zero for a crack size of zero, we believe that this is acceptable on two counts. First, the FCP represents an estimate of the POD for very small crack sizes. Second, the FCP in Phase II (and as described later, in Phase III) was small to begin with, implying that fitting the regression model to the data, either with or without the inclusion of FCP, produced similar results.

Using the approaches described previously, the average POD as a function of crack COD, using data from all five participating teams, is shown in Figure 5-3. The false call data used in these fits are those produced by “blank” material, not surface features. In the figure, the top two plots display the estimated curves surrounded by 95% confidence bounds, while the bottom two plots overlay the fit on the data from Phase II. The plots on the left represent the POD for cracks in ceramic specimens, while the top and bottom plots on the right are from the stainless steel specimens.

The bottom two “diagnostic” plots can be used to evaluate how well the regression model fits the data. Each data point in these plots represents a single crack that has been inspected by five teams. The POD point associated with each crack is surrounded by 95% bounds. If the curve fits the data, it should be within most of these bounds. This is seen to be the case.

Similar analyses, with length as the independent (explanatory) variable, are shown in Figure 5-4. While the POD curves as a function of length appear to be reasonable, the bottom diagnostic plots in Figure 5-4 indicate that there are two long cracks in the stainless steel specimens (one of length 55 mm [2.165 in.] and the other [not shown in plot] of length 80 mm [3.15 in.]) that were not easily detected. The longer of these cracks is on the weld edge (or weld toe) and was difficult to detect for all the participants.

Additional analysis (described further in Appendix F) indicated that, for the data from Phase II, COD appeared to have a stronger effect on detection performance for cracks in the stainless steel specimens. While this seems intuitive (cracks with larger COD may present a greater contrast with the background), crack length may be more relevant for component safety calculations.

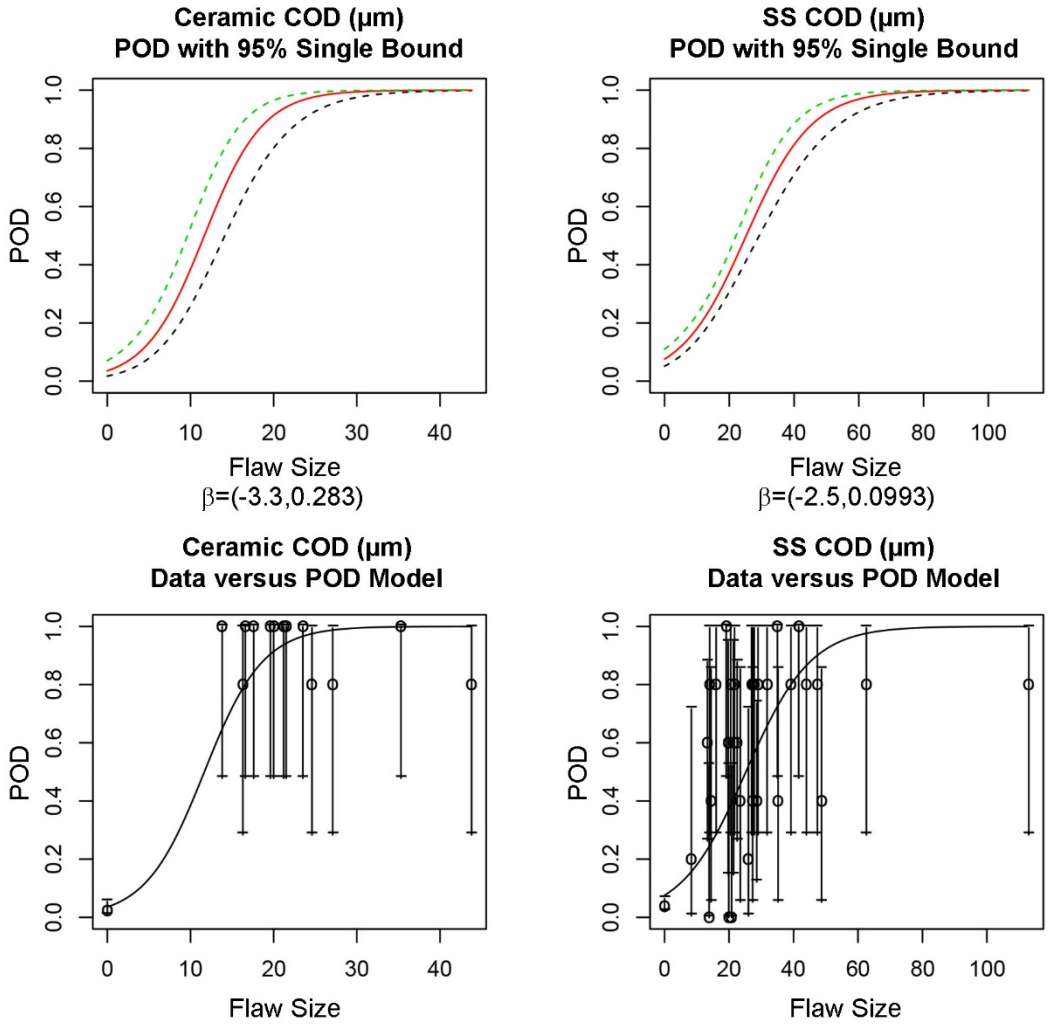


Figure 5-3 POD vs. COD for Ceramic (*left*) and Stainless Steel (*right*) Specimens

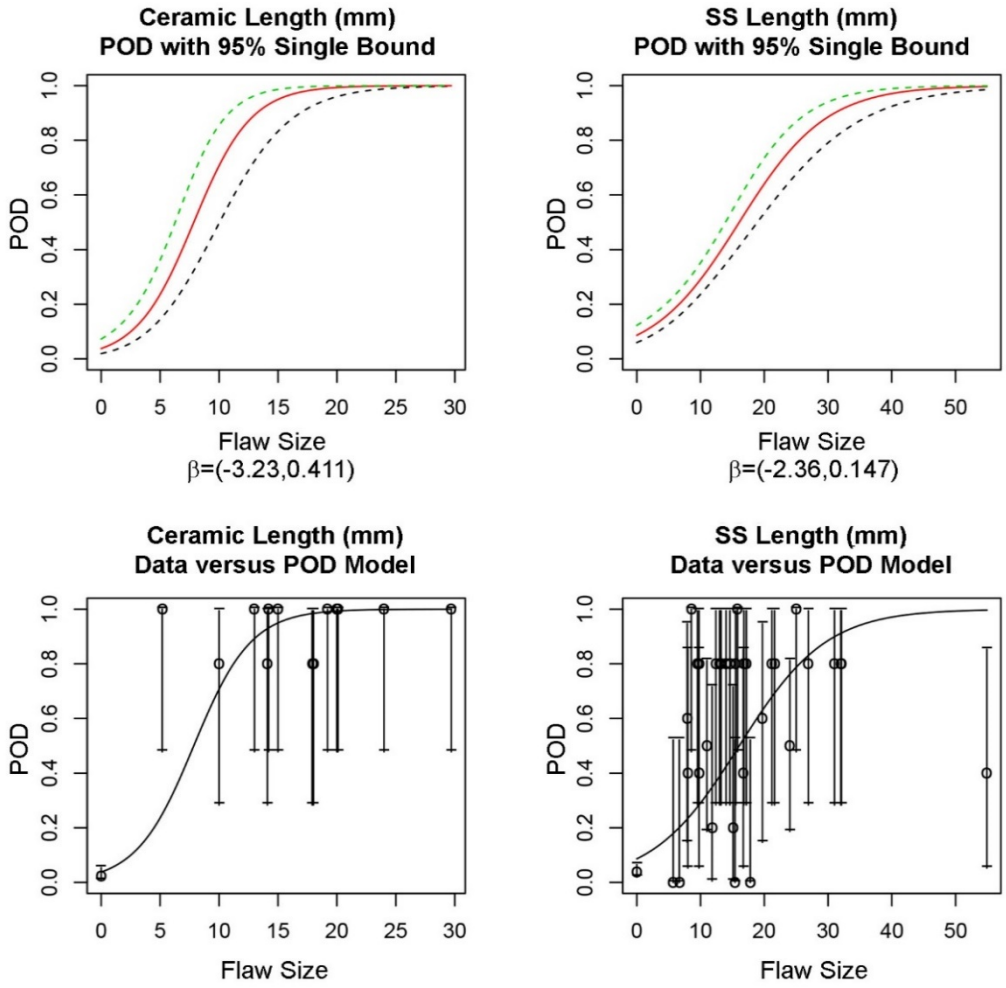


Figure 5-4 POD vs. Length Using All Detections and False Calls in Clean Material. 80 mm (3.15 in.) long crack, along the weld toe in a stainless steel specimen, is not shown.

The somewhat lower influence of length on detection performance raises questions about how cracks are detected and identified visually, and whether RVT is an appropriate technique for detecting cracks if it relies primarily on COD instead of length which, along with through-wall depth, are the important variables for safety calculations. Based on the analysis to date, and an examination of the data used in grading as well as the video data, several points were identified. First and foremost, there appears to be a link between COD and length [for example, Figure 5-1, or Wåle (2006)] for naturally occurring cracks. As a consequence, it is likely that longer and deeper cracks will have a greater COD, affording greater detection performance of the RVT method. Second, an examination of the video data also showed that in many instances the shorter cracks were visible in the recorded data, leading to a conclusion that perhaps improving the analysis methodology may improve the ability to detect shorter cracks. Finally, analyses conducted using alternative independent metrics (such as total area exposed by the crack opening or using both COD and length as independent variables) did not indicate conclusively whether length and COD alone or in combination provided a better explanation for the detection performance.

Interpreting the POD curves is often challenging. A lower POD for smaller cracks does not necessarily mean an inability to detect the crack. Rather, it means that as the crack gets smaller, reliable detection is more challenging. In this context, reliable detection refers to: if the same component is inspected multiple times by one or more teams, how often would a crack be detected?

Note that inspection reliability is a function of several parameters (including instrument factors and human factors). Depending on how some of these other factors vary from inspection to inspection, the overall inspection reliability can be affected. A case in point is team EQZH in Phase II. It is not clear why this team had a lower POD than the other teams. However, this reduced detection performance indicated an overall increased difficulty (when averaged over all teams) in detecting smaller cracks, represented by a shift in the POD curves towards larger cracks. The result is a reduction in inspection reliability for smaller cracks (small COD and/or shorter cracks).

A breakdown of the POD curves by inspection team, along with additional analyses, is given in Appendix E.

5.5.1.3 Detection Performance in Stainless Steel Specimens

The results presented in the previous sections indicated difficulties in detecting cracks in the stainless steel specimens. Assuming that these difficulties are because of the specimens themselves and not other factors, the data from the stainless steel specimens were further analyzed to determine if a relationship existed between POD and other variables such as crack orientation and location. For location, the analysis was conducted based on whether the crack was in one of four regions: (1) HAZ which includes up to 50 mm (2 in.) base metal[1]; (2) weld region with the weld crown ground flush; (3) weld region with the weld crown NOT ground flush; and (4) in a surface feature in the HAZ. Given concerns raised around detecting cracks in the weld toe, an additional analysis step examined whether evidence existed in the data indicating that cracks in the weld toe may be more difficult to detect.

Tables 5-9 through 5-11 present the resulting POD for these analyses. Table 5-9 shows a relationship between orientation and POD. However, orientation and crack location are related to each other, with transverse cracks only present in the weld region and circumferential cracks present outside the weld region. It seems likely that the relationship present in Table 5-9 is actually due to crack location as illustrated in Table 5-10; POD in the HAZ is higher than the other locations.

[1]This designation for the HAZ was used only for the purposes of this assessment. In typical inservice examinations, the HAZ is not expected to extend 50 mm (2 in.) into the base material on welds.

Table 5-9 POD vs. Flaw Orientation in Stainless Steel Specimens

Team	Transverse	Circumferential
ALYJ	0.69±0.12	0.82±0.06
BMXR	0.44±0.12	0.69±0.07
CIWN	0.69±0.12	0.69±0.07
DOYP	0.62±0.12	0.64±0.08
EQZH	0.00±0.04	0.21±0.07

Table 5-10 POD in Stainless Steel Specimens of Flaws in Different Locations: in Ground Weld, in Not Ground Weld, in HAZ, in Surface Feature

Team	Ground Weld	HAZ	Surface Feature	Not Ground Weld
ALYJ	0.67±0.20	0.83±0.08	0.84±0.09	0.57±0.19
BMXR	0.33±0.20	0.74±0.09	0.58±0.11	0.57±0.19
CIWN	0.67±0.20	0.74±0.09	0.63±0.11	0.71±0.18
DOYP	0.50±0.20	0.74±0.09	0.53±0.11	0.71±0.18
EQZH	0.00±0.11	0.17±0.08	0.21±0.10	0.00±0.09

It was noticed that two very large circumferential cracks placed on the weld edge were not detected at all, and this motivated the analysis presented in Table 5-11. From Table 5-11, we see that cracks located right at the weld crown edge are indeed difficult to detect.

Table 5-11 POD in Stainless Steel Specimens of Circumferential Flaws on Weld Toe

Team	Not on Weld Toe	On Weld Toe
ALYJ	0.88±0.06	0.57±0.19
BMXR	0.75±0.08	0.43±0.19
CIWN	0.75±0.08	0.43±0.19
DOYP	0.72±0.08	0.29±0.18
EQZH	0.25±0.08	0.00±0.09

5.6 Outcomes from Phase II

Phase II pointed to a generally lower than expected POD for cracks in components with surface texture resembling the stainless steel specimens. While the analyses of cracks in these specimens (Section 5.5.1.3) indicates that the reduction in POD may be attributable to the presence of other features on the specimen surface, such as scratches or proximity to the weld toe, the relative lack of influence of crack length on POD was also unexpected. Further, examination of the raw video data indicated that several cracks were visible in the data, indicating a potential negative influence of inadequate discrimination methodologies (to discriminate cracks from other surface features) on the result.

Discussions within the round-robin administration team (PNNL and EPRI) pointed to a few additional factors that may have played a role in the poorer than expected performance of fielded RVT inspection procedures. These factors include:

- Inspection Team Composition – Field inspection teams typically have at least two analysts (primary and secondary), where the secondary analyst is available to review and possibly correct findings reported by the primary analyst. The administration team questioned whether the presence of a secondary analyst in Phase II would have resulted in improved POD.
- Practice Specimens – Phase II had a limited number of specimens for the participating teams to use for practice. Given the experience from Phase I (taking the test multiple times resulted in improved detection), additional specimens that could be used by the teams to familiarize themselves with the specimens/cracks was proposed as a key need.
- Realistic Flaws – While most of the cracks in the stainless steel specimens were deemed to be realistic (i.e., mimicked cracks found in the field), a few were identified that may not have been realistic. These were generally grouped into two categories, extremely small cracks that were generally less than 8 mm (0.315 in.) long *and* a COD less than 10 microns (0.0004 in.), and weld toe cracks that did not deviate from the weld toe. The latter, in particular, was a point of extensive discussion as cracks found in the field in the weld toe region have tended to deviate occasionally into the HAZ. Such tortuosity may be beneficial as it affords opportunities to detect the crack without interference from the weld itself. However, the Phase II results also pointed to the difficulty of detecting cracks in the field that entirely follow the weld toe, which may partially explain their absence from the list of field cracks that have been experienced. A related issue was the use of notches in the specimen set. While no notches were present in the stainless steel specimens, Phase I results indicated improved ability to detect notches. The ceramic specimens, on the other hand, contained only notches, and the relatively robust detection POD (close to 100% for most teams) raised questions about the added value of using this specific combination of specimen and crack type in subsequent tests.
- Analysis Guidance – Each of the participating teams used their field procedures for Phase II; however, several cracks not detected by the teams were clearly visible in the raw video data. In several instances, the inspection team was found to have examined the indication (using multiple angles and by changing lighting) before concluding that the indication was not reportable as a crack. This finding implied that supplemental guidance on analysis procedures for discriminating between cracks and non-cracks may be helpful.
- Lighting – Phase II did not use auxiliary lighting; instead, the inspection teams were restricted to use on-camera lights only. The use of additional lights may improve the ability to discriminate between cracks and surface features.

Phase III of this study addressed many of these factors through a redesigned protocol and specimen set.

6 PHASE III OVERVIEW AND RESULTS

6.1 Overview

The Phase II RV-RRT activity pointed to the need for improved procedures for enhancing detectability of cracks and the need to answer several open questions. Addressing these open issues led to Phase III of this assessment. Given that many of the open questions dealt with cracks in stainless steel specimens, ceramic specimens were not used for Phase III. All of the stainless specimens from Phase II were used and were augmented with additional specimens fabricated at both PNNL and EPRI. However, very small cracks and many of the weld toe cracks in the Phase II specimens were judged to be not representative of field experience. Therefore these cracks were not required to be detected and did not count for or against performance metrics in Phase III.

The additional specimens included a diverse set of cracks, including cracks in the weld toe region, many of which reproduced operational experience by occasionally deviating from the weld toe into the HAZ. Additional blank specimens were also fabricated, as were several new specimens for use as practice specimens by the inspection teams.

The protocol also was modified to allow a secondary analyst on each team, to allow for live video re-inspections of some specimens (based on need), and auxiliary lighting. These modifications, and the results of Phase III, are discussed in greater detail in this section.

6.2 Objectives

The goal of the RV-RRT-3 was to assess the performance of commercially applied examination procedures augmented with improvements for enhancing detectability of cracks with qualified personnel, and identify areas for future improvement if needed.

The RV-RRT-3 had the following specific objectives:

- Identify and quantitatively assess enhancements to remote visual examination techniques for detecting cracks in test specimens.
- Evaluate improvements (over Phase II) to commercially applied examination procedures for their effectiveness.
- Quantify the impact of secondary review of all recorded examination data.
- Assess the level of image degradation (if any) from live to recorded data.
- Quantify performance improvements in terms of POD and FCR and determine the effect that certain important factors have on POD. Important factors for Phase III include:
  - Crack opening displacement
  - Crack length
  - Crack detection in the presence of surface irregularities or blemishes.

In addition, Phase III conducted a limited assessment to quantify degradation in recorded video data from RVT inspections.

6.3 Experimental Design

Phase III was limited to the use of stainless steel specimens, given the findings from Phase II. The test matrix specimen and crack design for Phase III used all 29 stainless steel specimens and associated cracks and surface features from Phase II. In addition, cracks from Phase II that were removed from the test population were replaced with similar cracks in terms of length and COD. Fourteen new stainless steel specimens were designed and fabricated by EPRI and PNNL for this purpose; these specimens included cracks over the range of COD and lengths, as well as cracks located near the toe of the weld as replacement for cracks in similar locations that were removed from the test population.

Table 6-1 summarizes the as-built crack matrix table for all the stainless steel specimens used in Phase III. This table describes the distribution of different length and COD combinations for the different weld conditions and orientations of the cracks. Table 6-2 summarizes the number of test specimens with multiple cracks per test specimen. Note that the numbers listed in Tables 6-1 and 6-2 include six cracks that were subsequently (post-Phase III testing) removed from the test population for not being representative of cracking experienced in the field.

Table 6-1 As-Built Flaw Matrix Table for All Stainless Steel Specimens Used in Phase III

Material	Weld Crown	Flaw				# Flaws (built)
		Fabrication Type	Orientation	Length	Width	
Stainless Steel	As Welded	Fatigue	Circ	Small	Small	6
				Small	Medium	2
				Small	Large	3
				Medium	Small	5
				Medium	Medium	5
				Medium	Large	9
				Large	Small	1
				Large	Medium	6
				Large	Large	9
			Axial	Small	Small	3
				Small	Medium	1
				Large	Small	1
				Large	Medium	0
				Medium	Small	6
				Medium	Medium	3
Total (stainless steel, as welded)						60
Stainless Steel	Ground	Fatigue	Circ	Small	Medium	5
				Medium	Small	4
				Medium	Medium	3
				Medium	Large	1
				Large	Medium	0
			Axial	Small	Small	2
				Small	Medium	1
				Large	Medium	1
				Medium	Small	4
				Medium	Medium	0
Total (stainless steel, ground)						21

COD: small = 5 to 20 µm (0.2 to 0.8 thou.), medium = 21 to 40 µm (0.82 to 0.1.57 thou.), and large > 40 µm (0.1.57 thou.). Length: small ≤ 15 mm (0.6 in.), medium > 15 mm (0.6 in.) and < 26 mm (1.02 in.), and large ≥ 26 mm (1.02 in.).

Table 6-2 As-Built Test Specimen Matrix Summary Table

Material	Weld Crown	Surface Roughness	Number of Specimens	Flaw Distribution: # of Specimens with:					
				0 Flaws (15%)	1 Flaw (27%)	2 Flaws (27%)	3 Flaws (18%)	4 Flaws (9%)	5 Flaws (5%)
Stainless steel	As welded	Typical	31	7	5	9	4	5	1
	Ground	Typical	12	2	5	2	1	1	1

6.3.1 Specimens

Unlike Phase II, all Phase III test specimens were constructed out of stainless steel with a simulated weld crown running down the middle of the specimen, and of approximately the same dimensions. Twenty-nine of the 43 Phase III test specimens were previously used in the Phase II test. The test specimens were designed to include all the different cracks and surface features one would expect to see in the field.

Table 6-3 provides a summary of the Phase III test specimens. Note that about 25% of the welds are ground and each specimen contains about two surface features on average. Surface features include surface blemishes such as dents, grind marks, scratches, or burn marks. Specimens contain from zero to five cracks with most containing less than three cracks. The 75 cracks mentioned in the table are intentionally planned cracks. Eight “bonus” cracks were unintentionally produced during the specimen fabrication process and were not necessarily representative of cracks in the field. These bonus cracks were not required to be detected and did not count for or against performance metrics.

Table 6-3 Summary of Test Specimens Employed in Phase III

# of Specimens Used in Phase III	43
Specimens with Ground Welds	12
Specimens with Unground Welds	31
Type of Specimens Used	Stainless steel
Average Length of Specimen	280 mm (11 in.)
Average Length of Blank Material in Specimens	146 mm (5.75 in.)
Total Cracks in Specimens	75
Specimens with No Cracks	9
Specimens with 1 Crack	10
Specimens with 2 Cracks	13
Specimens with 3 Cracks	6
Specimens with 4 Cracks	4
Specimens with 5 Cracks	1
Surface Features in Specimens	98

Tables 6-4 and 6-5 provide an overview of the orientation and location of cracks in the Phase III specimens. Most cracks are circumferential in orientation (parallel to the weld). The locations identified in Table 6-4 are on the weld edge (toe), in the weld, and in a surface feature. Table 6-4 also provides some information about surface features in the last row. As in Phase II, axial cracks were only located in the weld, while circumferential cracks were almost always located outside the weld. There are many more circumferential cracks in the HAZ and in surface features than there are axial cracks.

Table 6-4 Summary of Flaws and Surface Features

Type	Orientation		On Edge		In Weld		In Surface Feature	
	Axial	Circum.	False	True	False	True	False	True
Bonus	7	1	7	1	1	7	5	3
Crack	20	55	63	12	54	21	52	23
Surface Feature	28	44	70	2	55	17	72	0

Table 6-5 Flaw Count by Flaw Location and Flaw Orientation

	Axial	Circumferential
GrWeld	5	0
HAZ	0	36
InSF	5	18
NGWeld	10	1

GrWeld = ground weld crown
HAZ = heat-affected zone (includes base metal)
InSF = crack in surface feature
NGWeld = not ground weld crown (i.e., weld crown in as-welded condition)

6.3.2 Flaw Distributions

Figure 6-1 displays flaw size (COD vs. length) for the various types of cracks used in Phase III. As in Phase II, the distribution indicates some correlation between COD and length (chiefly associated with the extreme values). Determining which of these two factors influences the POD should still be possible if the effect is strong. For cracks in the field, we might expect a higher correlation between COD and length, based on other studies (Wåle 2006). Figure 6-1 also indicates that the size range for circumferential cracks is greater than that for axial cracks, likely because the length of the axial cracks was limited to the width of the weld. Table 6-6 provides a summary of the distribution of crack COD and length.

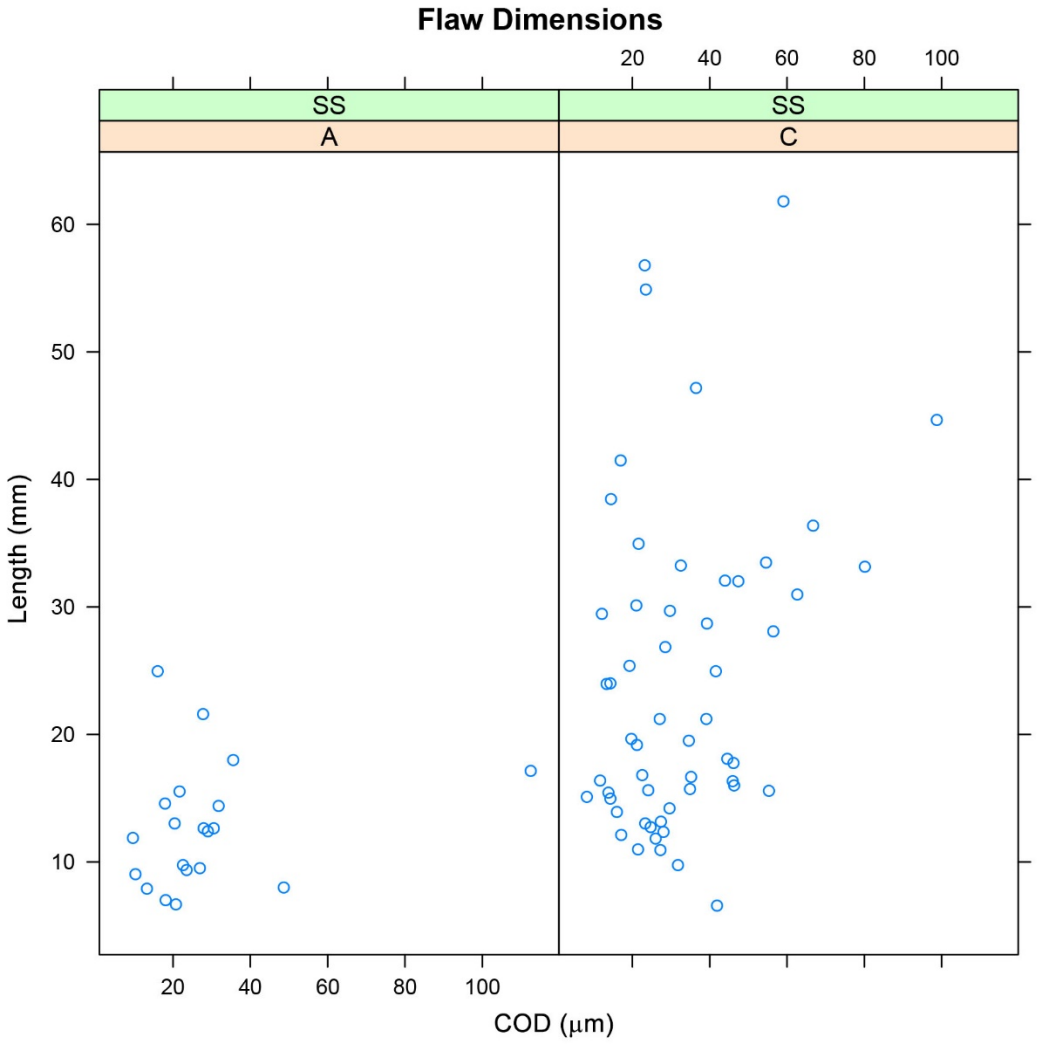


Figure 6-1 Plot of Flaw Dimensions Organized by Orientation (Axial, Circumferential)

Table 6-6 Quantiles of COD and Length for Flaws

	0%	25%	50%	75%	100%
COD Axial, µm*	9.58	18.06	23.01	29.39	112.51
COD Circumferential, µm*	8.28	21.09	28.08	42.96	98.76
Length Axial, mm (in.)	6.7 (0.26)	9.3 (0.37)	12.5 (0.49)	14.8 (0.58)	25.0 (0.98)
Length Circumferential, mm (in.)	6.6 (0.26)	15.3 (0.60)	19.7 (0.77)	31.5 (1.24)	61.8 (2.43)

*To convert microns to inches, multiply microns by 0.00004.

6.3.3 Test Methodology

The methodology for Phase III again used an RRT. Five teams from ISI service providers participated in a blind test where the true condition of the specimens was not revealed. The participating teams were asked to determine if a specimen contained a crack, and its approximate

location, orientation, and length. As in Phase II, their ability to accurately estimate the length of the crack was not evaluated and was used primarily to determine the extent of the crack for automated grading of crack detection.

As in Phase II, each team was only allowed to take the test once. Unlike Phase II, the test protocol (Appendix B) included two analysts, primary and secondary, in each team, was representative of typical field-inspection protocols, and was approved by the EPRI/Industry Remote VT Steering Committee. The test protocol allowed the primary analyst to evaluate live data and record indications that may be considered to be cracks. The primary analyst dispositioned these indications into one of three categories (relevant or cracks, non-relevant or not a crack, and relevant but needs re-inspection). All inspection data were recorded and the data along with the primary analysts' determination was provided to the secondary analyst. Using these, but without having the ability to directly question the primary analyst, the secondary analyst was asked to determine if an indication was a crack or not, and was allowed to add additional indications as cracks. In addition, the secondary analyst was allowed to select a subset of indications for re-inspection, where the questionable area on the specimen could be inspected again using the same camera system. The live data from this re-inspection were analyzed by the secondary analyst (in consultation with the primary analyst if necessary) and a determination made as to whether the area contained a crack or not. The recorded data from the initial inspection and re-inspection were also made available to the test administrators (PNNL and EPRI) for use in answering any questions regarding detection during the grading process. As in previous phases, the identifications of the specimen and participating teams were anonymized.

The inspection was performed underwater, with ambient lighting being limited by use of a black-out tent and camera motion controlled using a motor-controlled scanning bridge. Teams were allowed to use auxiliary lighting in addition to available lighting on the cameras (which were supplied by the participating teams). The inspection teams were given considerable flexibility on determining when to deploy auxiliary lighting, and at least one team used auxiliary lighting for most of the inspection. The mechanics of specimen movement in and out of the tank remained the same as in Phase II.

Prior to each test, teams were asked to perform a resolution check; all teams used the ASME character standard. Each team was also asked to perform any other necessary pre-inspection steps as required by their procedures prior to taking the test.

Assessment of the potential for image quality degradation in recorded data was performed by using the Air Force Resolution Target (DoD 1959). Each team was asked to image the Air Force Resolution Target. A staff member at EPRI was asked to identify the highest-resolution line pairs in the live data and subsequently performed the same analysis using the recorded data.

6.3.4 Data Description and Grading

Five teams participated in Phase III and inspected the 43 specimens described in the last section. All five teams inspected the same set of cracks, resulting in a balanced experimental design that simplified the comparison of team performance with averages. The same can be said for comparisons of performance on different categories of crack—an average calculated over each category described nominal inspector performance.

The Phase III crack evaluation procedure was more complicated than that used in Phase II. Two analysts participated in crack detection. The primary analyst performed the initial inspection, with the secondary analyst reviewing the results. The secondary analyst had the option to re-inspect

and alter the call as necessary. From the perspective of crack detection, there were three "stages" of the dispositioning process that could be evaluated. Each stage is associated with a step in the detection decision procedure being employed. The steps in crack detection that were tracked in the inspection reports were:

- **Recording Step** – An indication found and recorded on the data sheet; this step was performed by the primary analyst.
- **Primary Disposition Step** – The primary analyst usually classified the recorded indication as "Yes" or "Re-inspect." If indication was classified as "Yes," we concluded that the primary analyst had called a crack at this location. In this case, if there was a crack, the primary detected the crack; if not, the primary made a false call. This was called either a "primary" detection or a "primary" false call.

  Occasionally the primary analyst was observed classifying the recorded indication as "No" (not a crack). If there was a crack at this location, the primary analyst incurred a missed detection; if not, the primary analyst correctly evaluated the indication and determined that it was due to some other factor (such as a surface feature).

  Indications classified as "Re-inspect" are reviewed by the secondary analyst to obtain a final dispositioning of the indication (see below).
- **Final Disposition Step** – The secondary analyst reviewed all data and re-evaluated all indications. A fraction of the indications (including some of those marked "Re-inspect" by the primary analyst) were re-inspected and dispositioned as "Yes" or "No" using the live data from the re-inspection process. At the end of this process, the secondary analyst performed the final disposition of the crack. If this was a "Yes," and the indication was associated with a crack, this was a detection. If the indication was not associated with a crack, it was a false call. This was called a detection/false call, or for more clarity, a final detection/false call.

  Note that the protocol for Phase III restricted the number of locations that could be re-inspected by a team. This was done to ensure that the testing for each team was completed within the time available for conducting the test. A consequence of this decision was that some of the indications marked "R" (for re-inspect) could not be re-inspected. In these cases, the final dispositioning remained as "R."

Given this multistep decision process, the following probabilities were defined:

- **POR**: Probability of Recording an indication by the primary analyst. This is the equivalent as the probability of detection during Phase II.
- **PODP**: Probability of Detection by Primary. This requires the primary to have categorized the associated indication as a crack.
- **PODF**: Probability of Detection Final (by Secondary). This requires the secondary analyst to have categorized the associated indication as a crack.

The three probabilities described above were computed through the use of a "grading unit" of material. Thus, for example, POR was calculated by

$$\text{POR} = \frac{\text{\# GUs intersecting with a recorded indication}}{\text{Total \# GUs}}. \qquad (6\text{-}1)$$

Using these definitions, it was now possible to define probabilities associated with SFs and blank grading units. These are false call probabilities, but to emphasize their relationship to the three levels of detection defined above, the following notation was used:

- **POR(crack), PODP(crack), PODF(crack)**: Detection probabilities associated with cracked grading units.
- **POR(blank), PODP(blank), PODF(blank)**: Detection probabilities associated with blank grading units. These are the same as false call probabilities at each of the three stages of the decision procedure.
- **POR(SF), PODP(SF), PODF(SF)**: Detection probabilities associated with SF grading units. These describe another type of false call probability. SF false calls describe false call performance when the inspected surface is not in a pristine state.

The notation emphasizes the relationship between a particular POD and its associated FCP. The associated POD and FCP are calculated exactly the same except that one uses cracked grading units, while the other uses “blank” or “surface feature” grading units.

6.3.5 Grading Tolerance

Figure 6-2 displays the relationship between grading tolerance and POR. Originally, as in Phase II, we used a symmetric grading tolerance that was the same in both the transverse (X) and circumferential (Y) directions. However, the test setup enabled the teams to provide higher precision location information along the circumferential direction as opposed to the transverse direction. Therefore, in Phase III we used asymmetric grading tolerances, where the tolerance in the transverse direction was one and a half times that in the circumferential direction. Figure 6-2 displays the POD vs. the tolerance in the circumferential direction, and indicates that a tolerance of 10 mm (0.39 in.) in this direction (and 15 mm [0.6 in.] in the transverse direction) may be sufficient to account for minor errors by the inspection teams in locating cracks. These tolerances were used for the grading and analysis performed for Phase III.

As in Phase II, a manual review of the automated grading results was necessary to correct for any large errors in the location of reported indications. In addition, about 60 reported indications (total over all inspection teams) were assigned by the automated grading procedure to more than one grading unit. This was a consequence of the larger grading tolerance, where indications located between two cracks result in more than one detection when using the automated grading procedure. Almost all of these multiple associations involved a crack and a surface feature, with one exception that was associated with two cracked grading units. All instances of such multiple detection were flagged by the automated detection algorithm and were manually corrected.

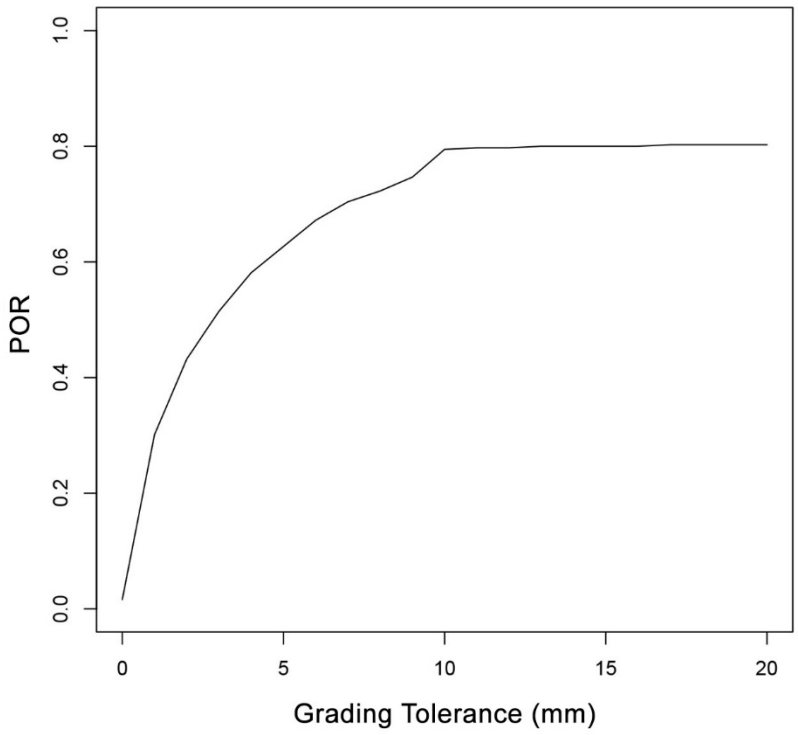


Figure 6-2 Plot of POR vs. Grading Tolerance in the Circumferential Direction. The grading tolerance in the transverse direction is 1.5 times that in the circumferential direction.

6.4 Phase III Results Summary

The previously described selection for grading tolerance, accompanied by a manual review and correction of any errors, was used for analysis. The analysis quantified procedure performance in terms of POD and determine the effect that certain important factors (COD, crack length, surface irregularities or other surface features) have on POD.

A summary of these results is provided below, along with a discussion on the potential implications and outcomes leading to Phase III. A detailed summary of the results is included in Appendix F.

6.4.1 General Findings from Phase III

6.4.1.1 Effect of Secondary Review

A significant difference in the protocol used for Phase III was the use of a team approach, where an independent (albeit informed) review of the inspection results was performed by a secondary analyst. The hypothesis was that the independent secondary review afforded opportunities to improve the POD by catching any indications that may have been missed.

Evaluation of this effect required contrasting the performance of the primary and secondary analysts. As described in Section 6.3.4, available data were categorized based on who dispositioned an indication and the result of this disposition. Possible outcomes of dispositioning by the analysts were: “Crack,” stating that the indication was a crack; “Not Crack,” stating that the indication examined was not a crack; and “Re-inspect,” stating that the indication needed to be re-examined.

Given the challenges with correctly dispositioning cracks located in surface features, and the increased probability of a false call in surface features (as seen from Phase II), we associated each indication with one of the three possible regions—crack, SF, or blank material. A correct disposition depends on the region the indication is in; an effective decision process will have placed all the Blank or SF indications in the “Not Crack” category, causing false calls to be zero. All cracks should fall into the “Crack” category, resulting in the highest possible detection probability.

Using this categorization, the inspection results from the primary and secondary analysts (aggregated over all teams) were analyzed, and the results shown in Table 6-7. Each row of this table lists the final dispositioning by a secondary analyst, when the initial dispositioning by the primary analyst was “Not Crack,” “Re-inspect,” or “Crack.”

Table 6-7 Disposition by Grading Unit Type

Initial Disposition	Final Disposition					
	GU = Blank		GU = Crack		GU = SF	
	Not Crack	Crack	Not Crack	Crack	Not Crack	Crack
Not Crack	2	0	1	5	3	1
Re-inspect	4	1	1	1	1	0
Crack	12	17	6	273	9	15

The first line in the table identifies those indications that the primary classified as “Not Crack.” To improve the results, the secondary analyst should have overruled the primary analyst for crack indications, but confirmed indications in both blank and SF material. The data show that the secondary did a fairly good job in this role; only two mistakes were made, one crack is left as “Not Crack” and one SF was incorrectly dispositioned as a crack.

In the second line where the primary explicitly requested a review by the secondary and possibly a re-inspection, we find that the secondary again made two incorrect calls (one blank dispositioned as a crack and one crack dispositioned as not a crack). The final row of Table 6-7 summarizes instances where the initial dispositioning was as a crack. In these cases, the secondary analysis resulted in 38 indications being incorrectly dispositioned. From these data, it appears that a critical evaluation by the secondary analyst of indications dispositioned as a crack may further detection performance.

Overall, however, this combined disposition procedure appears to be effective at reducing false calls. Most of the indications identified by the primary analyst as a crack are correctly identified (279 out of 287, or about 97%) after the final dispositioning. Of the 36 indications reported by the primary analyst in blank regions, only 50% are incorrectly identified as having a crack after final dispositioning by the secondary analyst. For SF material, the final disposition classifies about 55% (16 out of 29) as cracked. These results indicate that this procedure reduces the false call

probability by about 50%, at the expense of reducing POD by about 3%. From these results we see that there may be an opportunity to further improve performance by altering the decision procedure so that more indications are correctly identified as false calls.

Table 6-8 presents recording and detection statistics using all inspection teams. If the dispositioning procedure employed by the inspection teams was perfect, we would expect to see that indications reported as a crack are indeed cracks, and that no indications are reported in regions with surface features or in blank regions (i.e., POR(Crack)=POD(Crack), and POD(SF)=POD(blank)=0).

Table 6-8 Recording and Detection Statistics for Types of GUs

	POR	PODP	PODF
Crack	0.79±0.02	0.77±0.02	0.77±0.02
SF	0.13±0.02	0.12±0.01	0.10±0.01
Blank	0.05±0.01	0.04±0.01	0.03±0.01

One would classify the disposition procedure as ineffective when one could do as well by simply guessing the disposition of each indication. In this case, we would expect to see that the fraction of recorded indications that are classified as a crack is the same, regardless of whether the recorded indication is in a region with a crack, region with surface features, or a blank region. Mathematically, this corresponds to:

$$\frac{\text{POD(Crack)}}{\text{POR(Crack)}} = \frac{\text{POD(SF)}}{\text{POR(SF)}} = \frac{\text{POD(Blank)}}{\text{POR(Blank)}} \tag{6-2}$$

From the perspective of this criterion, the disposition procedure shows some effectiveness. For example, POD(Crack)/POR(Crack) = 97% while POD(SF)/POR(SF) = 77%, demonstrating that POD is reduced by 3% through the independent secondary review, while false calls (in SF) are reduced by 23% through this process. In blank material, false calls are reduced by about 40%. These numbers are in line with the raw totals listed in Table 6-7.

These results indicate that the secondary review's principal contribution to detection performance lies in the reduction of false calls. While the primary and secondary analysts were not allowed to communicate with each other during the test, the secondary review and decision-making is not completely independent, as the secondary analyst was provided with the initial dispositioning results and the raw video data for use in the review. It is not clear to what extent the fact that a secondary review was expected would have influenced the primary analyst's decision making. Phase III was not set up to extract this type of information. It is possible that the primary analysts used a more stringent criterion for their analyses, if they assumed that an independent secondary analysis that overturns many of the initial dispositioning may have negative consequences. On the other hand, it is possible that the primary analyst used a less stringent criterion, assuming that the secondary analysis would "fix" any mistakes. It is also possible that if the primary knew their decision was to be the final decision (as it turned out in most instances), they might have classified more indications as "not cracked," and thus reduced the FCP more dramatically.

The current exercise was not designed to produce detailed evaluations of teaming efforts (under multiple scenarios and options), and if deemed important, additional work in this area would need to be performed to determine the effect on the overall inspection reliability for RVT.

6.4.1.2 Summary of Detection and False Call Rates

The POD and FCP using a grading unit tolerance of 10 mm (0.39 in.) in the circumferential direction (and 15 mm [0.6 in.] in the transverse direction) are presented in Table 6-9. The results in Table 6-9 are separated by inspection team, and the FCP values are presented for blank material as well as for regions containing surface features. It should be emphasized that POD in this section refers to the "final" POD (i.e., PODF as determined by the secondary inspector's disposition).

Table 6-9 POD, FCP, and FCRs by Vendor

Inspection Team Code	POD	FCP in Surface Features	FCP in Blank Material	FCR (false calls/meter)
ARLW	0.72±0.05	0.11±0.03	0.01±0.01	0.159±0.193
DCSI	0.72±0.05	0.06±0.03	0.06±0.02	1.120±0.42
NBIE	0.85±0.04	0.13±0.03	0.02±0.01	0.319±0.249
TUQZ	0.80±0.05	0.10±0.03	0.03±0.02	0.638±0.331
YPJH	0.77±0.05	0.09±0.03	0.02±0.01	0.319±0.249
All Teams	0.77±0.02	0.10±0.01	0.03±0.01	0.510±0.128

The FCP in Blank Material represents a false call probability for a unit of material of 50 mm (2 in.) in length.

From this table, teams NBIE and TUQZ appear to achieve the highest POD while also achieving relatively low FCP. The "All Teams" row presents average performance over all participating teams. On average, POD is seen to be about 77% with an FCR of 0.51 indications/ meter.

Comparing these values with those obtained from Phase II (Table 5-8) on stainless steel specimens, we see that the POD appears to have improved in Phase III while the FCR has appreciably decreased. These changes could be due to a number of reasons, including more opportunities to practice prior to taking the test, improvements to inspection procedures, and the use of a secondary review of the results from the primary analyst.

As described in Section 6.4.1.1, the secondary review appears to have played a major role in the reduction of the FCP. Phase III was not set up to separate out the effects of procedure improvements and additional practice. However, the combined effects of inspection procedure improvements and additional opportunities to practice on representative specimens should manifest as an improvement in POD for cracks regardless of size and location. Below, we discuss additional analyses to assess the variations in POD as a function of crack size and location.

6.4.1.3 POD Analyses

As in Phase II, POD curves were computed using a regression curve fit to the data, and again, a logistic regression model that incorporated the FCP was used. The data indicated that COD, and not length, had a stronger influence on the POD (Appendix G). However, for completeness, POD as a function of both COD and length were computed.

While the results presented in this section utilize the logistic regression model, other regression models were also evaluated using Phase III data. This was done primarily to address concerns

that the models being used in Phase II and Phase III may not have been adequate to model the available data across the full range of COD and length values. Some details of this evaluation are presented in Appendix F, while a more comprehensive assessment (including unconventional regression models such as the Box-Cox model) is presented in Appendix G.

These assessments indicated the following:

- The logistic regression model used was adequate to fit the data over the entire range of COD and length values.
- While the logistic regression model may under-represent the POD for large cracks, this effect is small and has no appreciable effect on the crack COD or length that can be detected with an 80% POD at a 95% confidence level.
- Alternative models that force the POD to go through zero for small cracks may be capable of also representing the POD over the entire range of crack COD and lengths; however, such models cannot take into account the FCP.
- The goodness of fit of both types of models is similar for the data from Phase III.
- Alternative models (such as the Box-Cox model) may be capable of higher fidelity representation of the POD at both the low and high end, while incorporating the FCP into the calculation. However, the difference in estimated POD for the Phase III data was small, with no appreciable difference in the crack COD or length that can be detected with an 80% POD at a 95% confidence level.

As a result, using the logistic regression model, the average POD curve as a function of COD, along with 95% confidence bounds and diagnostic plots, are presented in Figure 6-3 and may be compared to Phase II results. Similar results, as a function of crack length, are shown in Figure 6-4.

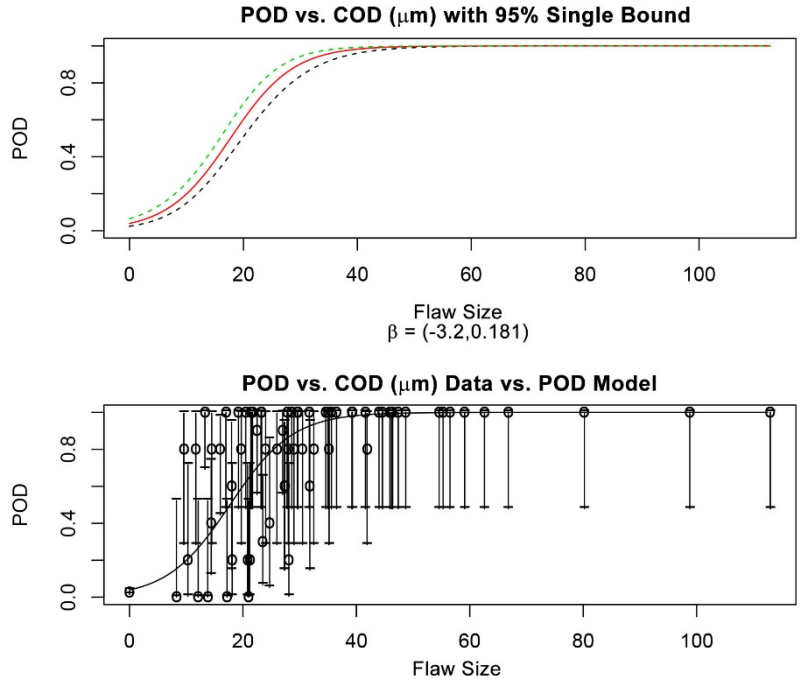


Figure 6-3 POD vs. COD

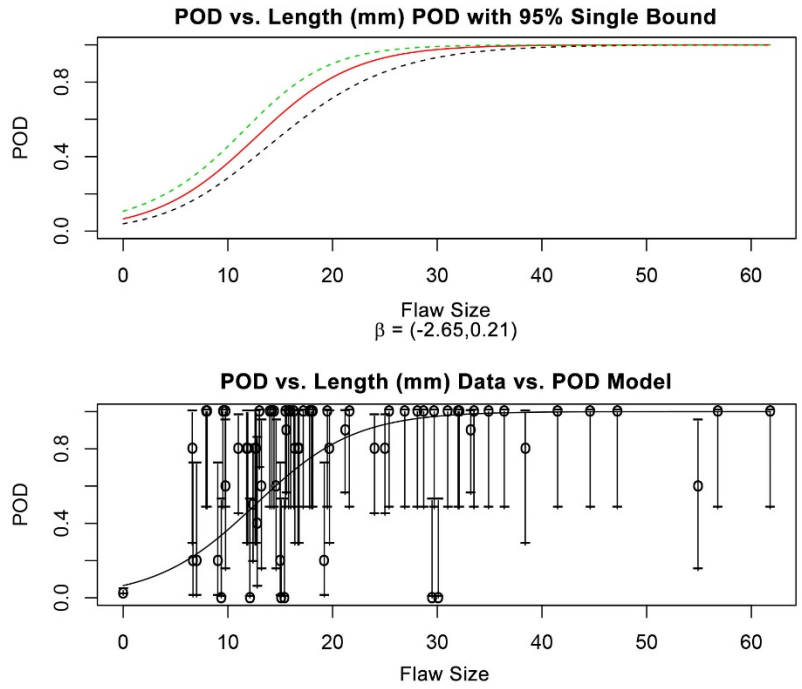


Figure 6-4 POD vs. Flaw Length

The diagnostic plots presented in Figures 6-3 and 6-4 show the data vs. the curve fit. Each “data point” represents a crack, with detection achieved by the five vendors. Each data point is surrounded by 95% confidence bounds and a point that does not fit would be indicated by a value whose bounds do not intersect the fitted curve. There are no such points (cracks) in the COD

curve fit, but there are such cracks in the length fit. For the length fit, note the cracks at 8, 19, 30, and 55 mm (0.315, 0.75, 1.2, and 2.165 in.). It is these cracks that diminish the goodness of fit (GOF) for the POD vs. length model.

It is important to note that the "poor" fit of the length model is not due to any deficiency in the regression model form, but due to crack-to-crack variations in detectability that cannot be accounted for by the variable crack length. Therefore, this POD model provides the best description of POD vs. length **for this set of cracks**.

Analyses also indicated that in Phase III, individual teams had fairly similar POD curve performance (Appendix G). The outlier is vendor DCSI, with a higher FCP and lower POD for large cracks. Because the five teams are so similar in their performance, we can conclude the RVT inspection protocol is achieving consistent performance.

Finally, Tables 6-10 through 6-13 present the crack sizes associated with a POD of 80% (Tables 6-10 and 6-12) and a POD of 90% (Tables 6-11 and 6-13) as these values of POD are generally considered reasonable target values for evaluating acceptable performance (Berens 2000; Generazio 2008; Annis et al. 2013). The tables also contain upper and lower bounds for the crack size estimates. From these tables, we see that 80% POD is reached at a COD of about 25 microns (0.001 in.) and length of approximately 19 mm (0.75 in.), while the 90% POD is reached at a COD of about 28 microns (0.0011 in.) and length of approximately 23 mm (0.91 in.). The confidence bounds presented here reflect crack sizes for a POD of 80% or 90%, at confidence levels of 2.5% and 97.5%. Similar bounds may be computed to reflect crack sizes at confidence levels of 5% and 95%. From the data presented here, the crack sizes for a POD of 90% at a confidence level of 97.5%] referred to as the 90/97.5 crack size or $a_{90/97.5}$ (Berens 2000) are seen to be a COD of approximately 33.61 microns and a length of approximately 27.5 mm.

Table 6-10 Estimate of Crack Size (COD) Associated with 80% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, µm*	Flaw Size, µm*	Upper Bound ($a_{90/97.5}$), µm*
ARLW	23.9	27.4	32.6
DCSI	24.8	29.3	36.3
NBIE	19.0	22.4	27.2
TUQZ	20.1	23.5	28.5
YPJH	20.8	24.0	28.6
All	23.0	25.4	28.4

*To convert microns to inches, multiply microns by 0.00004.

Table 6-11 Estimate of Crack Size (COD) Associated with 90% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, μm*	Flaw Size, μm*	Upper Bound ($a_{90/97.5}$), μm*
ARLW	27.5	31.6	38.2
DCSI	29.7	35.2	44.0
NBIE	22.8	26.4	32.5
TUQZ	23.7	27.7	33.9
YPJH	24.0	27.7	33.4
All	27.0	29.9	33.6

*To convert microns to inches, multiply microns by 0.00004.

Table 6-12 Estimate of Crack Size (Length) Associated with 80% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, mm (in.)	Flaw Size, mm (in.)	Upper Bound ($a_{90/97.5}$), mm (in.)
ARLW	17.1 (0.67)	19.9 (0.78)	24.3 (0.96)
DCSI	17.4 (0.69)	21.0 (0.83)	27.0 (1.06)
NBIE	12.9 (0.51)	15.2 (0.60)	18.5 (0.73)
TUQZ	16.3 (0.64)	19.6 (0.77)	24.9 (0.98)
YPJH	17.1 (0.67)	20.4 (0.80)	25.7 (1.01)
All	16.7 (0.66)	19.2 (0.76)	22.7 (0.89)

Table 6-13 Estimate of Crack Size (Length) Associated with 80% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, mm (in.)	Flaw Size, mm (in.)	Upper Bound ($a_{90/97.5}$), mm (in.)
ARLW	20.0 (0.79)	23.4 (0.92)	29.0 (1.14)
DCSI	21.0 (0.83)	25.5 (1.00)	33.1 (1.30)
NBIE	15.3 (0.60)	18.0 (0.71)	22.2 (0.87)
TUQZ	19.7 (0.78)	23.8 (0.94)	30.7 (1.21)
YPJH	20.5 (0.81)	24.6 (0.97)	31.4 (1.23)
All	20.0 (0.79)	23.1 (0.91)	27.5 (1.08)

A comparison with Phase II indicates an overall improvement in the POD curves as a function of both length and COD. This is consistent with the observed improvement in the overall POD described in Section 6.4.1.1, and is likely due to one or more of the same factors described earlier.

6.4.1.4 Detection Performance as a Function of Other Variables

Phase II results had indicated difficulty in detecting cracking if there are other surface features nearby. These features could include scratches, grind marks, or other factors that may limit visibility of the crack, such as proximity to the weld toe.

To determine if the procedural and protocol changes in Phase III produced an observable change in any of these factors, these same factors were further evaluated using the data from Phase III. The evaluation is presented with categorical tables that display POD (as a percentage) and a standard deviation of the POD estimate. As in the analysis of Phase II data, the standard deviation estimates presented here are a function of the sample size (number of grading units). These tables do not consider crack size, and might therefore lead to incorrect conclusions if the crack sizes of the different categories of cracks differ greatly. These tables are meant only to provide an overview of the effect of these variables.

Table 6-14 presents POD by inspection team and crack orientation. Except for team NBIE, the POD for circumferential cracks is slightly greater than that for transverse (axial) cracks. In fact, if the results from all teams are combined (presented in the "All Teams" row of the table), we see that axial and circumferential POD differ by only 3 percentage points, which is within the standard deviation (shown in Tables 6-14 through 6-17 as error bars against each quantity). Note that the standard deviation when accounting for all teams is smaller than those computed for individual teams; this is largely due to the larger sample size when data from all teams is combined.

Table 6-14 POD (%) of Axial/Circumferential Flaws by Vendor

Inspection Team	Transverse	Circumferential
ARLW	65±11	75±6
DCSI	70±10	73±6
NBIE	95±6	82±5
TUQZ	75±10	82±5
YPJH	70±10	80±5
All Teams	75±4	78±2

Table 6-15 presents PODs for different crack locations. From the "All Teams" row, there is weak evidence that cracks in surface features are harder to detect than cracks at the other locations. There is no evidence that RVT on ground welds behaves any differently than on unground welds, or that detection in the HAZ is different than these other two locations.

Table 6-15 POD (%) of Flaws in Different Locations: in Ground Weld, in HAZ, in Surface Feature, and in Unground Weld

	Ground Weld	HAZ	In Surface Feature	Unground Weld
ARLW	80±19	75±7	65±10	73±14
DCSI	80±19	72±8	65±10	82±12
NBIE	100±12	83±6	78±9	100±6
TUQZ	80±19	83±6	70±10	91±10
YPJH	80±19	81±7	70±10	82±12
All Teams	84±8	79±3	70±4	85±5

Tables 6-14 and 6-15 are similar to tables presented in Phase II and can therefore be used to compare Phase II performance with Phase III. Such a comparison indicates an overall improvement in the detection of transverse cracks and a slight improvement in detection performance in the presence of surface features.

Tables 6-14 and 6-15 do not account for any relationships that might exist between crack orientation and locations. For example, all axial cracks are in weld crowns. Table 6-16 presents the effect that orientation and location jointly have on POD. These results indicate that transverse cracks are hardest to detect when in a surface feature (POD is reduced from about 84% to 48%). For circumferential cracks, HAZ and surface feature locations produce about the same POD, while cracks in the weld have a higher POD.

Table 6-16 POD (%) of Flaws in Different Orientations/Locations

	A: Ground Weld	A: In Surface Feature	A: Unground Weld	C: HAZ	C: In Surface Feature	C: Unground Weld
ARLW	80±19	40±22	70±15	75±7	72±11	100±43
DCSI	80±19	40±22	80±13	72±8	72±11	100±43
NBIE	100±12	80±19	100±7	83±6	78±10	100±43
TUQZ	80±19	40±22	90±11	83±6	78±10	100±43
YPJH	80±19	40±22	80±13	81±7	78±10	100±43
All Teams	84±8	48±10	84±5	79±3	76±5	100±12

Table 6-17 presents the effect of weld crown edge (toe) on POD. As seen from Phase II, circumferential cracks on the weld crown edge were somewhat difficult to detect. Analysis in Phase III indicates that this challenge was not addressed through the procedure and protocol changes that were implemented, and that cracks on the toe of the weld continue to present a challenge for visual examinations.

Table 6-17 POD (%) of Circumferential Flaws on Weld Toe

	Not on Weld Toe	On Weld Toe
ARLW	79±6	58±14
DCSI	79±6	50±14
NBIE	84±6	75±13
TUQZ	88±5	58±14
YPJH	86±5	58±14
All Teams	83±3	69±6

6.4.1.5 Recorded Data and Video Quality Degradation

Phase III included a resolution check step where the inspection teams were asked to record video of an Air Force Resolving Power Test target (Figure 6-5) (DoD 1959). Simultaneously, an independent analyst from EPRI was asked to review the live data and quantify the resolution in both the horizontal and vertical directions, by identifying the smallest line pair that could be resolved. The same analyst was then asked to review the recorded data from this test target and identify the smallest line pair that could be resolved.

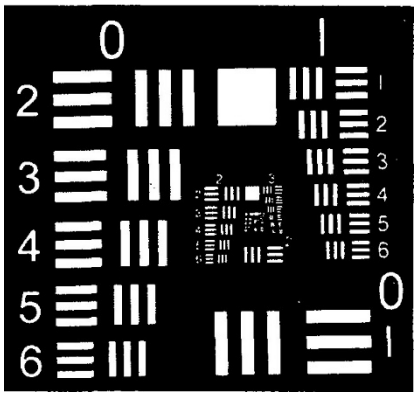


Figure 6-5 Sample Image of a 1951 U.S. Air Force Resolving Power Target (Cumblidge et al. 2007)

The resulting data from this subjective test indicated that no significant degradation in recorded video quality existed. This result contradicted the observations from previous phases of this research and the reason for this apparent disconnect is not entirely clear. Possible explanations include:

- The observed data from previous phases indicated that the degradation in recorded data quality depended on the surface texture of the specimen, with the greatest degradation observed with the ceramic specimens. The resolution target used, on the other hand, was fabricated on clear glass allowing for an excellent test of resolution but without the accompanying complexity of specimen surface texture.
- The amount of data quality degradation (subjectively evaluated) in previous phases appeared to be a function of the camera systems used. This may point to the possibility of changing the level of compression used within the recording software. It is not clear how the systems (and

software) that were used in Phase III compared with those used in previous phases, and whether any changes that were made to the Phase III instruments played a role in the observed data quality.

We should note that the data quality from the test specimens in Phase III did not show any clear evidence of degradation. This is also in line with the observed data quality from stainless steel specimens in Phase I and II, although the recorded data quality from ceramic specimens in Phases I and II showed a clear degradation when compared to the live data.

All of this points to the need for a more rigorous assessment of this issue in the future, through the use of resolution targets that are etched (or otherwise mounted) onto specimens that display the necessary surface texture. Even without these studies, the end users of these systems should be aware of possible data quality issues with the recorded data and account for these issues in the inspection plan and procedures.

7 DISCUSSION

The three phases of research under this assessment led to several insights on the use of RVT for IVVI. Discussions with participants also provided context and identified areas where the findings may be leveraged in a relatively short time frame. These insights and opportunities are described below.

7.1 Comparison of Phase II and Phase III Results

Phase III generally showed improved results, in terms of POD and FCP, when compared to Phase II. The crack parameters for reliable detection improved (smaller CODs, smaller lengths were detected at the 80% POD level) in Phase III when compared to Phase II. This appears to be the result of a combination of procedure improvements and the use of additional practice specimens with multiple flaws. The use of a secondary analyst, while helpful in reducing false calls, did not help improve the POD.

7.2 Reliability of RVT for IVVI

Insights around the reliability of RVT for IVVI may be broadly categorized into the following areas:

7.2.1 General Comments and Observations

- In a few cases, indications in the close vicinity of flaws were called by an inspection team, but it was not always clear whether the flaw had been detected, or some other surface feature next to the flaw was detected. Where possible, manual review of the raw video feeds was used to assist in resolving this question. In instances where this was not apparent even with manual review, the operator was given the benefit of the doubt in this study by assuming that the crack had been detected. From the data and discussions between PNNL, EPRI, and some of the inspection teams, it was not clear whether this is an issue in actual field examinations.
- For the procedures and specimens used in this study, COD appeared to have a larger effect on detection performance than crack length.
- In general, the data from the three phases of study indicate generally reliable detection (80% POD at 95% confidence) for flaws with COD above approximately 25 micron (0.001 in.). Similarly, flaws with length above approximately 20 mm (0.8 in.) appeared to allow for reliable detection using the improved RVT procedures. Note that flaw length and COD are somewhat correlated, and therefore the influence of one on the other needs to be fully accounted for when assessing performance. It is also worth noting that, while smaller COD is usually also associated with shorter cracks, limited studies have shown that the relationship between COD and length can be poor for some forms of degradation (Cumblidge et al. 2004). Indeed, operationally, it is possible that crack closure (due to external or internal loads) may occur for deep internal surface cracks in thick walled piping during shut down periods (Wåle and Ekstrom 1995); such conditions may influence the COD and length relationship and influence the ability to detect cracking.
- FCP cannot be ignored when evaluating detection performance. In many important situations, such as clean vs. scratched material, the difference is not in POD, but FCP.
- Acceptable or critical flaw sizes are still unknown and will depend on the specific component and plant.

- One interesting observation was that documentation errors existed in the data. These generally took the form of transcription errors or other errors in documenting the location of cracks by the inspection teams. It is not clear whether this may have been the result of the test protocol and documentation data sheets that were used, and how the observed documentation error rates might compare with field practice. The general consensus, based on discussions between PNNL and EPRI, was that these types of errors in the field may be identified and fixed during the several iterations of review during typical field examinations. As a result, this line of enquiry was not pursued further within this study.

7.2.2 Specimens and Flaws

- The data seemed to indicate that surface textures similar to those in the stainless steel specimens were more difficult to inspect than surface textures represented by the ceramic specimens in Phase II. However, a complicating factor here was the use of different flaw fabrication methods for the stainless steel which used fatigue cracking and ceramic specimens with laser-cut notches. The limited data from Phase I indicated that laser notches were somewhat easier to detect than fatigue cracks, given the same type of specimen surface texture. This could be due to the inherent limitations in fabricating laser notches with small COD as well as the heightened contrast between notches and any background texture. Together, these findings point to the need to use realistic flaw fabrication methods for training and evaluation of RVT performance.
- Surface features, such as scratches or grind marks, appeared to have the major effect of increasing FCP but not necessarily decreasing POD. This may incorrectly indicate that evaluation methods for discriminating between cracks and non-crack surface features are potentially sufficient. However, the overall POD for cracks in these regions was lower than that for cracks in clean (i.e., no surface features) regions, indicating the need for improved guidance in examination procedures for discriminating between cracks and non-crack surface features. This issue is further discussed in the section on examination procedures.
- Surface features that mask the crack or otherwise make it difficult to view and the ability to apply lighting at an appropriate angle will significantly challenge RVT detection performance. This is particularly true for cracks at or near the weld toe. While supplementary guidance, practice, and auxiliary lighting seems to help improve detection in many of these cases, the weld-toe region appears to be particularly challenging and may be a limiting case for the use of RVT.
- Based on feedback from some of the inspection teams, the specimens used in Phase II and Phase III appeared to be representative of field conditions, in both surface texture and the use of realistic cracks. These teams expressed an interest in using similar specimens for training purposes.

7.2.3 Cameras and Instrumentation

- The three phases of this research study indicated that overall, different camera systems appeared to be capable of similar detection performance. Clearly, there are differences between the different camera systems that were used by the participating teams. However, many of these are related to camera features such as pan-tilt-zoom vs. tube type, color vs. black and white, standard resolution vs. high definition, etc. Other factors in selection of camera systems are ease of component access and radiation tolerance.
- One concern was the potential for degradation of video quality when dealing with compressed video recordings, based on Phase I, and to a limited extent, Phase II data. The use of

compression is generally an advantage when dealing with the storage and transmittal of large amounts of video. Several video compression standards (such as MPEG-4) exist and appear to be used by most RVT camera vendors. However, compression algorithms may result in the loss of some types of information, especially in cases where the texture content of images in the video is somewhat low. In the present study, this issue appeared to be particularly problematic with the ceramic specimens and was likely a result of the specific texture used for the specimen surface. Data collected during Phase III using a resolution standard did not indicate a significant difference between live and recorded data quality. However, the compression settings for recording appear to be tunable (this was not specifically examined during this study), and it is possible that these settings may change from inspection to inspection in the field. As a result, end users of this technology may need to take appropriate action to ensure that a significant change in recorded data quality does not occur from inspection to inspection, and enables robust analysis from the recorded data.

- Auxiliary lighting (for the most part, diffuse lighting) appeared to be used extensively for evaluation and re-inspections and in some instances during the scanning process. While the improved detection performance in Phase III cannot solely be attributed to the use of auxiliary lighting, based on a review of selected video data, it is likely that the use of auxiliary lighting was beneficial overall.

7.2.4 Examination Procedures

- Field inspection procedures, based on a review of selected procedures, appeared to provide considerable flexibility to the inspector in terms of choice of equipment and its use for crack detection. In general, such flexibility is beneficial as it allows the inspector to adapt to conditions “on the ground,” and does not inadvertently contribute to reduced detection performance by being overly prescriptive. However, the extreme flexibility provided by current procedures ensures that the inspection result is dependent on the skill and experience of the inspector.

  In such cases, past experience with NDE in both nuclear and non-nuclear industries has led to mechanisms for ensuring that different inspectors have some minimum skill level that is considered acceptable. Several approaches to addressing this question may be possible. The data from this assessment point to one possible method, so-called “guided practice,” for addressing this issue that uses realistic specimens and flaws, along with supplemental guidance on evaluating and dispositioning indications, for training and level-setting of inspection skills. A more robust approach might be through performance demonstration, as defined in ASME Section V Article 14 (ASME 2015b), for RVT systems including equipment, procedures, and personnel.

- The use of secondary analysis in typical field-applied inspection procedures appears to help reduce false calls. In the tests conducted within this study, secondary analysis did not appear to change the POD but was able to confirm that the right component was examined and that inspection coverage of the component was sufficient. Given that the protocol included the ability to perform re-examinations as in the field (albeit, in Phase III, on a limited set of indications given time constraints), the potential for secondary analysis to boost POD appears to be limited. Note that Phase III of this study examined a specific scenario wherein the secondary analyst conducted a review of the primary analyst’s results but was not allowed to consult with the primary analyst during the primary or secondary review. Data from this study does not reveal whether joint analysis of the inspection data would help in improving POD. Such studies will need to be conducted using human factors approaches that the present study was not designed to perform.

7.2.5 Other Factors (including Human Factors)

- While the different teams appeared to have similar detection performances, one team had a significantly lower detection rate in Phase II. The reason for this difference was not apparent in the data but could have been due to multiple factors, including perceived time pressures, stringent detection thresholds, or any of a number of other factors. As a result, the detection performance of this one team should not be considered to be typical for in-field inspection, even though one or more such factors may influence typical in-field inspection as well. Instead, the results from this assessment, specifically the generally similar performance in Phase III, should be considered a baseline (and possibly an upper bound) from which performance for typical field inspections can vary as other factors are incorporated.

  As discussed in the previous section, guided practice using representative specimens, flaws and surface features, with truth information provided to the inspection team appears to be useful in ensuring a minimum level of performance. This finding appeared to be reinforced by the results from the training effect seen from multiple tests in Phase I and in Phase III where the use of guided practice specimens, in combination with other procedure improvements, appeared to improve the performance from Phase II.

- The value of supplementary guidance on evaluating potential indications was not clear from the data. Much of the supplementary guidance appeared to repeat information that was apparent to the inspectors, and the inspection team generally did not refer to the guidance during the Phase III test. As a result, a possible value of the supplemental guidance may have been as a tool for reminding inspectors of best practices.

7.3 Expected Consistency of RVT Inspections in the Field

A major question is the applicability of the results from this assessment to RVT inspections in the field to quantify POD and FCR in typical field inspections. Extrapolating the results to a typical field inspection is a challenge and will require augmenting the data with quantities such as the surface feature density in field components and FCRs in the field. There is also an open question with respect to the effect that deployment systems (rope/pole or robotic systems) will have on the POD. Given these open questions, some of which are difficult to quantify in a laboratory setup, the results reported in this document should be considered as a mechanism to help identify factors that may influence the POD and FCR in a field examination. Given that the assessment indicates improvements in POD can be obtained through specific actions, such as better training, these actions should be considered for implementation with the expectation they will lead to improvements in field inspection performance.

A related question is the required POD and FCR for field inspections. Answering this question requires addressing the unresolved issue of acceptance criteria for flaws. As discussed in previous sections, such information is plant and component specific, and often proprietary to the plant. Limited studies seem to suggest that for certain components such as core support structures, the acceptable flaw dimensions are large enough to be reliably detected using EVT-1 (Nickell and Rashid 2001). However, the present assessment did not address this question and the results presented should be considered as a means to identify factors, such as the presence of surface features, crack adjacency to the weld toe, or other geometric/physical features that impact optimizing lighting and viewing angles, that may limit the ability to detect cracks in specific components. A better understanding of these factors can lead to potential improvements in RVT instrumentation as well as in the appropriate selection of other inspection methods (such as UT) that may provide the necessary sensitivity for specific components.

7.4 Unresolved Items

As discussed in Sections 3 and 4, several variables that may affect RVT performance were not explicitly controlled during the assessment. These include:

- Oxide build-up on internal components
- Thermal distortion of video images
- Water currents and clarity
- Radiation effects on camera video quality
- Limits on accessibility, viewing angle, and lighting
- Camera delivery systems
- Personnel qualification levels.

All of these factors are expected to impact RVT performance, but to different levels. Further, the impact of some of these factors may be limited by procedures used in the field; as a result, further studies on those factors are not considered a high priority at this time.

RVT performance relative to other NDE methods is also an unresolved item. This evaluation was intentionally deferred until after this assessment was completed as a quantitative evaluation of RVT in lieu of other NDE methods (such as UT) was not possible until these data were available.

Potential uses for VT-1 (and potentially VT-2 and VT-3 examinations) seem to be increasing, with proposed use of these methods to inspect spent nuclear fuel dry storage canisters, and for advanced reactors (liquid metal and high temperature gas reactors). As small modular reactors come on line, it is expected that VT in general (and RVT in particular) may also play a role in assuring the integrity of components. However, these newer applications appear to bring additional challenges with respect to types and location of cracking, access restrictions, and cracking precursors, that may or may not challenge existing instrumentation and procedures, and may require additional skills-development and capability assessments for RVT inspection teams. Proposed automated analysis techniques for RVT are also likely to become commonplace. Such techniques were discussed during the development and conduct of this assessment, but were ultimately not included as the technology was not deemed to be sufficiently mature.

These developments in RVT technology and anticipated challenges in applying VT to different systems point to the need for continued evaluation of the capability and effectiveness of the inspection technology. Given these developments, it is also likely that there may be a renewed push to use VT over other NDE techniques (and indeed, it may be the only option in some cases). As a result, it may be appropriate for future work to include studies that benchmark the performance of VT with respect to other NDE methods, targeted for specific components and cracking mechanisms.

8 SUMMARY AND CONCLUSIONS

Remote visual examination or RVT is a commonly used NDE method for inservice inspection of reactor internals to detect cracking and gross component failures. A major open question with regard to the use of RVT is the reliability of remote visual examination. This report described the results of an assessment conducted for the purpose of determining the reliability of RVT.

A series of RRTs were conducted to identify factors affecting RVT reliability and to quantify the POD as a function of these factors. These studies showed that COD is a major factor in the reliability of crack detection using commercially applied RVT procedures, with crack length being less impactful. Practically, the results imply that RVT detection is heavily dependent on the contrast produced by the crack opening for a given COD. Conversely, while crack length appears to positively correlate with detection probability, the correlation appears to be weak. This result, in turn, implies that crack detection using RVT is increasingly less reliable as the COD decreases, particularly below about 25 microns (0.001 in.). Note, however, that unreliable detection is not the same as no detection, it simply means that the probability that the crack will be detected every time is low.

The assessment also reinforced earlier findings regarding the importance of lighting in RVT detection, and appeared to reinforce other studies that find improved reliability when using multiple inspectors or independent analysts. This specific issue was not thoroughly studied, given that it would require a well-designed human factors analysis.

The results also point to the importance of training, especially with specimens that mimic the specimen conditions likely to be found in the field. Procedures that explicitly describe the decision process may be helpful as a reminder to the experienced analyst when it comes to discriminating between a reportable indication and one that is not reportable (i.e., is likely a surface feature).

Intuitively, RVT will be challenged when cracks are in the vicinity of other surface features such as scratches or weld ripples, or close to the toe of welds where shadowing and/or the presence of weld undercuts may complicate the ability to detect cracks. These hypotheses were supported by the results from Phases II and III and point to some limitations of RVT.

Based on the findings, and the limitations of the studies, the following recommendations are made (in no particular order of importance):

- RVT procedures should be updated to include additional details on performing the inspection and guidance for discriminating between cracks and non-cracks. While this information may be ingrained in experienced analysts, such information may be helpful as a reminder for all analysts. This is particularly important as, in many cases, the camera systems used in this assessment were observed to be capable of imaging the cracks. In such instances, a missed detection is almost certainly due to the decision processes for discriminating between cracking and non-cracking.
- Specimens that mimic the surface conditions and types of cracks likely to be encountered in the field should be used for training purposes prior to inspection teams performing field examinations.
- The limitations of RVT should be in the forefront when planning or analyzing data from an inspection. Consideration should be given to the use of alternate techniques for inspecting challenging areas such as weld toe regions.

- The applicability of RVT should be determined in close conjunction with the development of crack acceptance criteria specific to the components being inspected. In many cases, it is likely that large cracks can be tolerated, such as in the case of a core shroud in BWRs; in these cases, the reliability of RVT should be sufficient to detect these well before failure of the component. In other instances where much smaller cracks need to be detected, the specific circumstances associated with the component (environment, minimum detectable flaw size, impact of missed detection, etc.) need to be considered prior to the application of RVT.
- It is likely that the camera deployment systems used will affect the overall reliability. This needs to be better quantified.
- Advances in RVT technology, such as HD cameras and automated image analysis algorithms, should be evaluated to determine if they can help further improve the reliability of RVT.
- The condition of the surface (surface texture, patina, oxide, or other deposits) may be important in detection. As discussed earlier, it is likely that surface texture plays a role in the clarity of recorded data, and limited data were obtained to evaluate the effect of patina on detection. However, this assessment did not extensively evaluate these factors, nor did it evaluate the effects of deposits and the effectiveness of cleaning procedures. These factors need to be better quantified.
- While a review of the detection results by a secondary analyst appeared to be effective at reducing false calls, this research was not set up to thoroughly evaluate the possible benefits of teams of inspectors or analysts. A further evaluation of these factors using well-controlled and well-designed human-factors studies will be needed for better quantification of the benefits of inspection teaming efforts.

9 REFERENCES

78 FR 37885. June 24, 2013. "10 CFR Part 50, Approval of American Society of Mechanical Engineers' Code Cases." *Federal Register* 78(121):37885-37920. Nuclear Regulatory Commission, Washington, D.C.

Agresti A. 1990. *Categorical Data Analysis*, Volume 218 of Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, New Jersey.

Annis C, L Gandossi and O Martin. 2013. "Optimal Sample Size for Probability of Detection Curves." *Nuclear Engineering and Design* 262(Supplement C):98-105. DOI: 10.1016/j.nucengdes.2013.03.059.

ASME. 2010. *Code Case N-648-1 Alternative Requirements for Inner Radius Examinations of Class 1 Reactor Vessel Nozzles Section XI, Division 1*. American Society of Mechanical Engineers, New York. Approved September 7, 2001.

ASME. 2015a. "IWA-2211, VT-1 Examination; Section XI, Rules for Inservice Inspection of Nuclear Power Plant Components." In *ASME Boiler and Pressure Vessel Code*. American Society of Mechanical Engineers, New York.

ASME. 2015b. "Section V, Nondestructive Examination; Article 14, Examination System Qualification." In *ASME Boiler and Pressure Vessel Code – An International Code*. American Society of Mechanical Engineers, New York.

ASME. 2015c. "Table IWA-2211-1, Visual Examinations; Section XI Rules for Inservice Inspection of Nuclear Power Plant Components." In *ASME Boiler and Pressure Vessel Code*. American Society of Mechanical Engineers, New York.

ASME. 2015d. *Code Case N-619 Alternative Requirements for Nozzle Inner Radius Inspections for Class 1 Pressurizer and Steam Generator Nozzles Section XI, Division 1*. American Society of Mechanical Engineers, New York. Approved February 15, 1999.

ASME. 2015e. "Section XI Rules for Inservice Inspection of Nuclear Power Plant Components." In *ASME Boiler and Pressure Vessel Code*. American Society of Mechanical Engineers, New York.

ASME. 2017. "Table IWB-3512-1, Allowable Planar Flaws; Section XI Rules for Inservice Inspection of Nuclear Power Plant Components." In *ASME Boiler and Pressure Vessel Code*, p. 125. American Society of Mechanical Engineers, New York.

ASNT. 2016a. *Recommended Practice No. SNT-TC-1A: Personnel Qualification and Certification in Nondestructive Testing*. American Society for Nondestructive Testing, Columbus, Ohio.

ASNT. 2016b. *ASNT Standard for Qualification and Certification of Nondestructive Testing Personnel*. ASNT CP-189, American Society for Nondestructive Testing, Columbus, Ohio.

Berens AP. 2000. *Probability of Detection (POD) Analysis for the Advanced Retirement for Cause (RFC)/Engine Structural Integrity Program (ENSIP) Nondestructive Evaluation (NDE)*

System Development, Volume 1 - POD Analysis. AFRL-ML-WP-TR-2001-4010, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Chen F-C, MR Jahanshahi, R-T Wu and C Joffe. 2017. "A Texture-Based Video Processing Methodology Using Bayesian Data Fusion for Autonomous Crack Detection on Metallic Surfaces." *Computer-Aided Civil and Infrastructure Engineering* 32(4):271-287. DOI: 10.1111/mice.12256.

Crawley MJ. 2012. *The R Book, Second Edition*, Wiley.

Cumblidge SE, MT Anderson and SR Doctor. 2004. *An Assessment of Visual Testing*. NUREG/CR-6860, PNNL-14635, U.S. Nuclear Regulatory Commission, Washington, D.C. ADAMS Accession No. ML043630040.

Cumblidge SE, MT Anderson, SR Doctor, FA Simonen and AJ Elliot. 2007. *A Study of Remote Visual Methods to Detect Cracking in Reactor Components*. NUREG/CR-6943, PNNL-16472, U.S. Nuclear Regulatory Commission, Washington, D.C. ADAMS Accession No. ML073110060.

DoD. 1959. *Military Standard - Photographic Lenses*. MIL-STD-150A, U.S. Department of Defense (DoD), Washington, D.C. Available at http://www.dtic.mil/dtic/tr/fulltext/u2/a345623.pdf.

DoD. 2009. *Department of Defense Handbook - Nondestructive Evaluation System Reliability Assessment*. MIL-HDBK-1823A, U.S. Department of Defense (DoD), Washington, D.C. Available at http://www.statisticalengineering.com/mh1823/MIL-HDBK-1823A(2009).pdf.

EPRI. 2005. *BWR Vessel and Internals Project, Reactor Vessel Pressure Vessel and Internals Examination Guidelines*. TR-105696-R8 (BWRVIP-03 Rev. 8), Electric Power Research Institute (EPRI), Boiling Water Reactor Owners Group's Vessel and Internals Project, Palo Alto, California.

EPRI. 2015a. *Materials Reliability Program: Pressurized Water Reactor Internals Inspection and Evaluation Guidelines (MRP-227, Revision 1)*. Final Report 3002005349, Electric Power Research Institute (EPRI), Palo Alto, California.

EPRI. 2015b. *Materials Reliability Program: Inspection Standard for PWR Internals -- 2015 Update (MRP-228, Rev. 2)*. EPRI Report 3002005386, Electric Power Research Institute (EPRI), Palo Alto, California.

Forli O. 1995. *Guidelines for Replacing NDE Techniques with One Another*. NT TECHN REPORT 300, Det Norske Veritas, Finland. Nordtest Project Number 1159-94.

Generazio ER. 2008. "Directed Design of Experiments for Validating Probability of Detection Capability of NDE Systems (DOEPOD)." In *Proceedings of the 34th Annual Review of Progress in Quantitative Nondestructive Evaluation, Volume 27A*, pp. 1693-1700. July 22-27, 2007, Golden, Colorado. DOI 10.1063/1.2902640. American Institute of Physics, Melville, New York. Available at http://aip.scitation.org/doi/abs/10.1063/1.2902640.

Hothorn T and BS Everitt. 2006. *A Handbook of Statistical Analyses using R*, Chapman and Hall/CRC Press. pp. 89-108.

IAEA. 2013. *Training Guidelines in Non-Destructive Testing Techniques: Manual for Visual Testing at Level 2*. Training Course Series No. 54 (IAEA-TCS-54), International Atomic Energy Agency, Vienna, Austria.

Landrum J and G Selby. 2005. *Evaluation of Remote Visual Examination Methods*. Technical Update 1011625, Electric Power Research Institute, Palo Alto, California.

Liu Z, H Ukida, P Ramuhalli and K Niel, Eds. 2015. *Integrated Imaging and Vision Techniques for Industrial Inspection - Advances and Applications*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag, London.

Luk KH. 1993. *Boiling-Water Reactor Internals Aging Degradation Study. Phase 1*. NUREG/CR-5754; ORNL/TM-11876, Nuclear Regulatory Commission, Washington, D.C. ADAMS Accession No. ML040300570.

McCullagh P and JA Nelder. 1983. *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Chapman and Hall.

Moran TL, P Ramuhalli, AF Pardini, MT Anderson and SR Doctor. 2010. *Replacement of Radiography with Ultrasonics for the Nondestructive Inspection of Welds - Evaluation of Technical Gaps - An Interim Report*. PNNL-19086, Pacific Northwest National Laboratory, Richland, Washington. ADAMS Accession No. ML101031254.

Newman TS and AK Jain. 1995. "A Survey of Automated Visual Inspection." *Computer Vision and Image Understanding* 61(2):231-262. DOI: 10.1006/cviu.1995.1017.

Nickell RE and YR Rashid. 2001. "Technical Justification for ASME Code Section XI Crack Detection by Visual Examination." In *International Conference on Nuclear Engineering*. April 8-12, 2001, Nice, France. Available at http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/33/003/33003488.pdf.

NRC. 2011. *Addendum of Memorandum of Understanding between U.S. Nuclear Regulatory Commission and Electric Power Research Institute on Cooperative Nuclear Safety Research -- Memorandum of Understanding for Nondestructive Examination*. U.S. Nuclear Regulatory Commission, Washington, D.C. ADAMS Accession No. ML103080165.

NRC. 2014. *Regulatory Guide 1.147, Inservice Inspection Code Case Acceptability, ASME Section XI, Division 1*. Rev. 17, U.S. Nuclear Regulatory Commission (NRC), Washington, D.C. ADAMS Accession No. ML13330A6989.

Pascual FB. 2014. "Detection of Cracks and Corrosion for Automated Vessels Visual Inspection." http://srv.uib.es/wp-content/uploads/2014/09/mthesis_Bonnin2010.pdf.

Schmugge SJ, NR Nguyen, C Thao, J Lindbergy, R Grizziy, C Joffey and MC Shin. 2014. "Automatic Detection of Cracks During Power Plant Inspection." In *2014 3rd International Conference on Applied Robotics for the Power Industry (CARPI)*, pp. 1-5 Foz do Iguassu. DOI 10.1109/CARPI.2014.7030042.

Spencer FW. 1996. *Visual Inspection Research Project Report on Benchmark Inspections, Final Report*. DOT/FAA/AR-96/65, U.S. Department of Transportation, Federal Aviation Administration, Washington, D.C.

Wåle J and P Ekstrom. 1995. *Crack Characterisation for In-service Inspection Planning*. SAQ/FoU-Rapport 95/07, SAQ Kontroll AB, Stockholm, Sweden. Available at http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/27/028/27028075.pdf.

Wåle J. 2006. *Crack Characterisation for In-service Inspection Planning - An Update*. SKI Report 2006:24, Swedish Nuclear Power Inspectorate, Stockholm, Sweden. Available at http://www.stralsakerhetsmyndigheten.se/Global/Publikationer/Rapport/Sakerhet-vid-karnkraftverken/2006/SKI-Rapport-2006_24.pdf.

Ware AG, DK Morton, JD Page, ME Nitzel, SA Eide and T-Y Chang. 1999. "A Program for Risk Assessment Associated with IGSCC of BWR Vessel Internals." In *Pressure Vessels and Piping Conference*. August 1-8, 1999, Boston, Massachusetts. Idaho National Engineering and Environmental Lab., Idaho Falls, ID (US). INEEL/CON-98-01263. Available at http://www.osti.gov/scitech/servlets/purl/8079.

APPENDIX A PHASE II PROTOCOL

A.1 Summary

The Remote Visual Round Robin Test (RV-RRT) comprises a study of the performance of remote visual non-destructive examination (NDE) techniques currently used for In-Vessel Visual Inspections (IVVI). The Round Robin inspections are intended to provide an assessment of commercially applied procedures using blind inspections.

The project will focus on inspections in the welded region of large components (using flat plates with simulated welds). The study will assess crack detection and discrimination, where discrimination means distinguishing between cracks and non-relevant surface features.

The project time schedule for completion of the round robin testing is first quarter of 2013.

This document describes the scope, prerequisites, organizations, and rules for the RV-RRT.

A.2 General

A.2.1 Scope of Remote Visual (RV) Round Robin Test (RRT)

The purpose of this PROTOCOL document is to describe the scope, prerequisites, organizations, and rules for RV-RRT.

The NRC, in cooperation with EPRI, is conducting an evaluation of the reliability of remote visual testing (EVT-1) currently used for IVVI. A limited round robin test was conducted during Phase I of this evaluation. The results of the Phase I test, along with a subsequent parametric study, were used to design a more extensive Phase II round robin test (designated as the RV-RRT in this document).

The RV-RRT comprises a study of the effectiveness of remote visual non-destructive examination (NDE) techniques currently used for IVVI. The Round Robin testing is intended to provide an assessment of commercially applied examination procedures using a blind testing methodology. The goal is to assess the performance of currently used procedures with qualified personnel. The results from the RV-RRT will be used to identify areas for future improvement, if needed.

The project will focus on remote visual inspections in welded regions, using flat plates with simulated weld crowns. The study will assess crack detection and discrimination, where discrimination means distinguishing between cracks and non-relevant surface features.

The RV-RRT will be carried out as a coordinated effort between PNNL and EPRI, with participating inspection teams from the United States (U.S.) with direct field experience with in-vessel visual inspection (IVVI) in nuclear facilities.

A.2.2 RV-RRT Objectives

The goal of the Remote Visual Round Robin Test is to assess the performance of commercially applied examination procedures with qualified personnel and identify areas for future improvement, if needed.

The Remote Visual Round Robin Test has the following objective:

- Identify and quantitatively assess remote visual examination techniques for detecting and characterizing flaws in test specimens.
  - Evaluate commercially applied inspection procedures for their effectiveness
    - Quantify procedure performance in terms of probability of detection (POD) and determine the effect that certain important factors have on POD. Important factors for Phase II include:
      - Crack opening dimension/displacement (COD)
      - Crack length
      - Crack detection in the presence of surface irregularities or blemishes

A.2.2.1 Expected Outcomes

The RV-RRT data are expected to provide a better overall understanding of the performance of commercially applied remote visual examination procedures and the critical factors that affect the performance.

In addition, from the RV-RRT data, be able to calculate the following:

- POD curves for each participating inspection team as a function of flaw size (COD and length).
- If the results from more than one team can be grouped (for instance, if the examination procedures utilized allow such grouping), POD curves will be created for these groups.
- POD curves for the different types of test specimens present in test.
- Identification of significant differences in POD related to important variables:
  - Examination procedure, test specimen type, flaw type, and orientation.
  - Evaluation of false call probability (FCP).

A.2.2.2 Limitations

The test specimens used in this round robin (and described in Section A.2.6) represent welds similar to that found in some reactor internal components. The specimens represent two specific colors (patina) in reactor internals (natural stainless steel and reddish tints) and may not be representative of the diverse color variations of internals surfaces. The effect of other configurations and surface conditions on test results is beyond the scope of this round robin. Finally, the RV-RRT is not designed to assess certain variables that may impact detection performance in remote visual testing. These are:

- Lighting options
- Oxide build-up on internal components
- Thermal distortion
- Water currents and clarity
- Radiation effects on camera video quality
- Limited accessibility (component configurations and camera size)

- Monitors and camera systems
- Camera delivery systems
- Personnel qualification levels
- The angle of view limits for Code VT-1 exams

A.2.3 RV-RRT Project Organization

NRC Program Manager:
Wallace E. Norris
U.S. Nuclear Regulatory Commission
RES
Washington, DC 20555-0001 USA

Industry Steering Committee Chair:
Chuck Wirtz
First Energy
USA

RV-RRT Project Contact Persons:
Dr. Pradeep Ramuhalli, PNNL
P.O. Box 999, MSIN K5-26
Richland, WA 99352 USA
Phone: 509-375-2763
Email: pradeep.ramuhalli@pnnl.gov

Michael T. Anderson
Scientist/Engineer
P.O. Box 999, MSIN K5-26
Richland, WA 99352 USA
Phone: 509-375-2523
Email: michael.anderson@pnnl.gov

Jeff Landrum, EPRI
1300 West WT Harris Blvd
Charlotte, NC 28262
Phone: 704-595-2553
Email: jlandrum@epri.com

John Lindberg, EPRI
Program Manager – NDE Innovation
1300 West WT Harris Blvd.
Charlotte, NC 28262
Phone: 704-595-2625
Email: jlindberg@epri.com

Industry Steering Committee:

- Chuck Wirtz, Chair Ad-Hoc Committee – BWRVIP, First Energy
- Tim Wells – MRP IIG Chair, Southern Nuclear

- Mark Huting – NDE IC Chair, Xcel Energy
- Marc Brooks – NDE IC, Detroit Edison
- Harry L. Smith – APC 4, Exelon
- Rich Ciemiewicz – BWRVIP, Exelon
- Dan Nowakowski – MRP ITG, NextEra
- Tony Oliveri – APC 4, Public Service Electric and Gas
- Jeff Landrum – EPRI
- Jack Spanner – EPRI
- John Lindberg – EPRI
- Greg Selby – EPRI

Test Administrators:

- PNNL – Dr. Pradeep Ramuhalli
- EPRI – Jeff Landrum
- EPRI – Chris Joffe
- EPRI – John Lindberg
- EPRI – Jonathan Buttram

A.2.4 Inspection Teams

The following inspection companies/teams are participating in the remote visual round-robin test.

Company	Team	Comments/Notes
Companies: Contact persons:	AREVA Team 1	
Companies: Contact persons:	Westinghouse Team 1	
Companies: Contact persons:	WesDyne Team 1	
Companies: Contact persons:	GE Team 1	
Companies: Contact persons:	IHI Southwest Team 1	

A.2.5 V-RRT Schedule

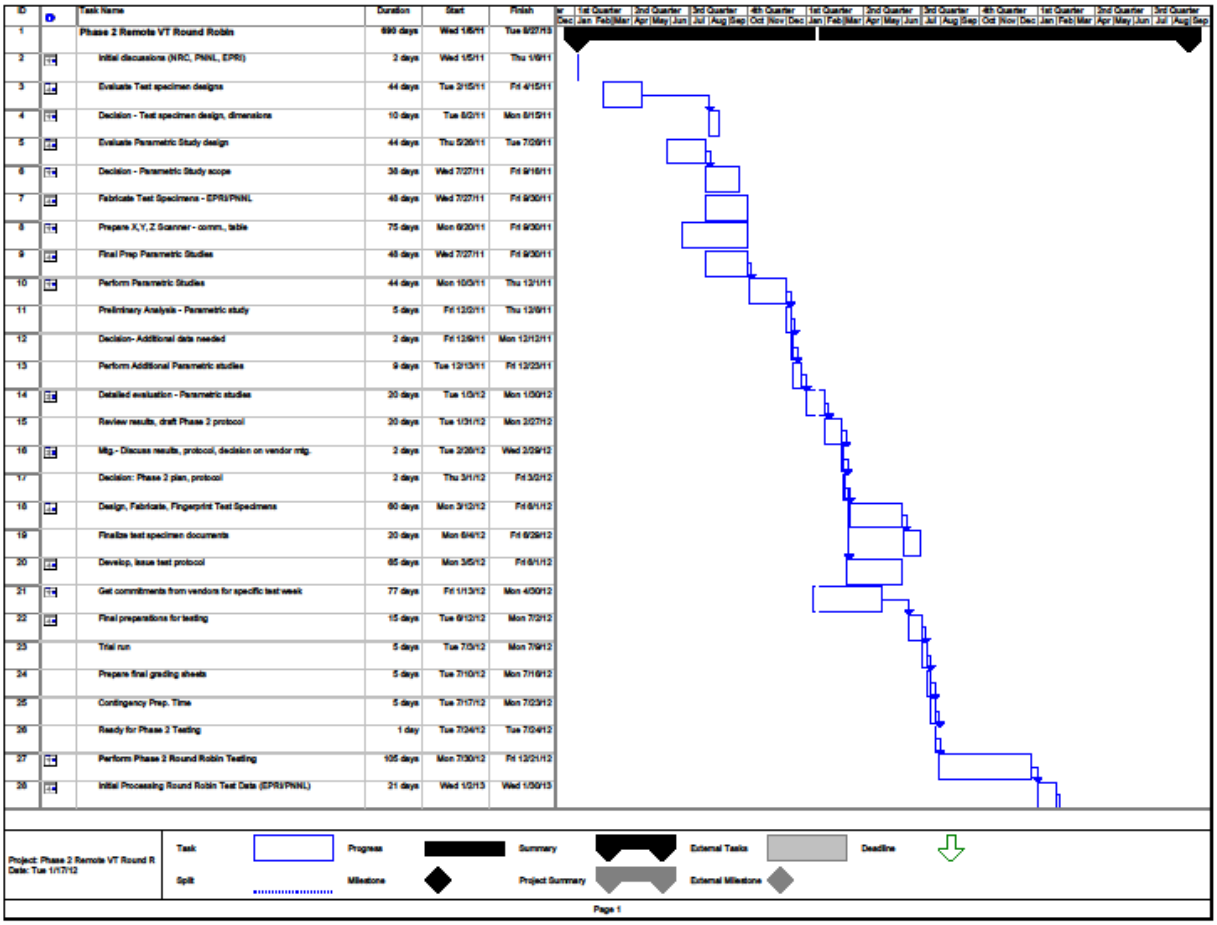


Figure A-1 Preliminary Schedule for RV-RRT (Updated: January 2012)

A.2.6 RV-RRT Specimen Information

A.2.6.1 Test Objects

Category	Typical Weldment	Comments/Notes
Ceramic Flat Plates	Butt weld	Simulated welds. Specimens have surface features such as tooling marks and scratches.
Stainless Steel Flat Specimens	Butt weld	Simulated welds. Specimens have surface features such as undercut, grind marks, tooling marks, and scratches.

The RV-RRT will consist of a sufficient number of cracked and blank grading units to enable calculation of POD and associated confidence bounds.

The specimens will be rectangular plates (dimension varies) of stainless steel or ceramic material with a simulated weld crown. The specimens will be clean, but not shiny, for the inspections. No additional cleaning of the specimens will be allowed.

A measuring scale will be etched, taped, or otherwise mounted on the specimen to assist the inspector in recording the location of any detected cracks.

Each sample will be labeled using a unique identifier (ALIAS name) that will be visible to the inspector.

Note: Unless there are mitigating circumstances, as approved by the test administrator, participating teams must examine all test specimens provided.

A.2.6.2 Inspection Region

The region of interest will be the base materials adjacent to, and including, the simulated weld on both the upstream and downstream sides. The inspection region will be identified for each specimen.

A.2.6.3 Defect Specification

The expected degradation mechanisms in internals components are stress corrosion cracking and fatigue cracking. The test specimens will have real and simulated cracks with defect orientation being transverse or longitudinal to the welding direction.

Discrimination between cracking defects and non-relevant surface features (such as scratch marks) shall be part of the evaluation process. Flaw length estimates are not necessary, except to the extent needed to correctly locate flaws relative to the weld/specimen location markers.

A.2.7 Data Security

All information concerning the blind tests is considered to be confidential and shall therefore be dealt with as such. Specifically, all parties participating in the RV-RRT, and/or in evaluating the results from the RV-RRT, shall not release or discuss any data, results, papers or data media, or any other information, to anyone not authorized for that type of information without prior approval from the NRC and the Industry Ad-Hoc Committee.

Authorized personnel are those who participate in the RV-RRT and the subsequent data analysis (U.S. Nuclear Regulatory Commission (NRC), Pacific Northwest National Laboratory (PNNL), Electric Power Research Institute (EPRI), Ad-hoc Committee Members, and the Test administrators).

Summary results from the RV-RRT will be published after the analysis of the results is complete. The identity of the inspection teams shall be anonymous throughout the testing period and afterward in published reports and documents.

All personnel in the RV-RRT Project Team, test administrators, and inspection teams conducting the RV-RRT must comply with this protocol.

The following restrictions shall be applied:

- All papers and information, including scrap papers and data media must be handed over to test administrators and those who are responsible for the RV-RRT activities. No unauthorized person may remove such information from the test facilities.

- During the RV-RRT, no Internet connections, electronic devices, or wireless devices (cell phones, iPhones, etc.) are allowed.
- Copying of data or data transformation is not allowed without prior permission from authorized personnel (test administrators).
- Removable data storage media (such as memory sticks) are not allowed in the test facilities, except for those accepted by test administrators to be used during the inspection.
- The RV-RRT data shall not be recorded onto media (such as computer disk drives) brought on-site by the inspection team. The test administrator shall supply storage media for recording the RV-RRT data.
- Neither the test administrators nor the participants in the RV-RRT may discuss defects or results with other participants in the RV-RRT.

Test specimens, test results, and teams will be assigned ALIAS names for recording data and results. The ALIAS names are produced by PNNL and EPRI, and are designed to conceal the identity of the items listed.

Only authorized personnel will have the knowledge of the real team names, test results, or test specimens used.

- All data generated in the RV-RRT will be provided to the test administrators in digital form.
- All test administrators must have a proven independence and impartiality status with respect to the participating teams.
- All test administrators need to have an approved back up in case they become sick or job conflicts prevent them from being able to perform their duties.

A.3 Document Review

A.3.1 Introduction

Each test administrator will review the Examination Procedures (EP) developed by the participating inspection teams for the RV-RRT. This requirement is put in place to facilitate review and analysis of the results of RV-RRT, and to quickly get answers to questions that arise during the testing, data analysis and review process.

Note: It is not the role of test administrators to make any comment about the procedures used by the participants.

The EP documents are to have a unique name, include an edition/revision identifier, are dated, and provide the name of the author, reviewer, and approver. They are to be made available in electronic format. This will help to ensure that test administrators can receive counsel from other test administrators and that right decisions can be made to resolve issues.

Examination Procedures should be sent to the test administrator sufficiently in advance of the scheduled test period. Test administrators will review these documents and write a summary document called Examination Procedure Summary (see Attachment A1), for review by PNNL and EPRI. This document will summarize the techniques for detection and flaw discrimination. The purpose of this document is to facilitate review and assessment of data (during data analysis)

without having to read all examination procedures in detail. The Procedure Summary shall not contain any proprietary information in the examination procedures.

A.3.2 Examination Procedure

Examination Procedures should contain all relevant information regarding the preparation, performance and reporting of remote visual inspection. The examination procedure should describe **what to do** and **how to do it**. It should be clearly stated in the scope of the examination procedure within which limits it is valid (for example, material type, surface finish, etc.).

The purpose of the RV-RRT is to assess commercial procedures that can be applied without any technique changes. However, some adaptations could be necessary in order to accommodate/fit the test specimen geometries and the test setup.

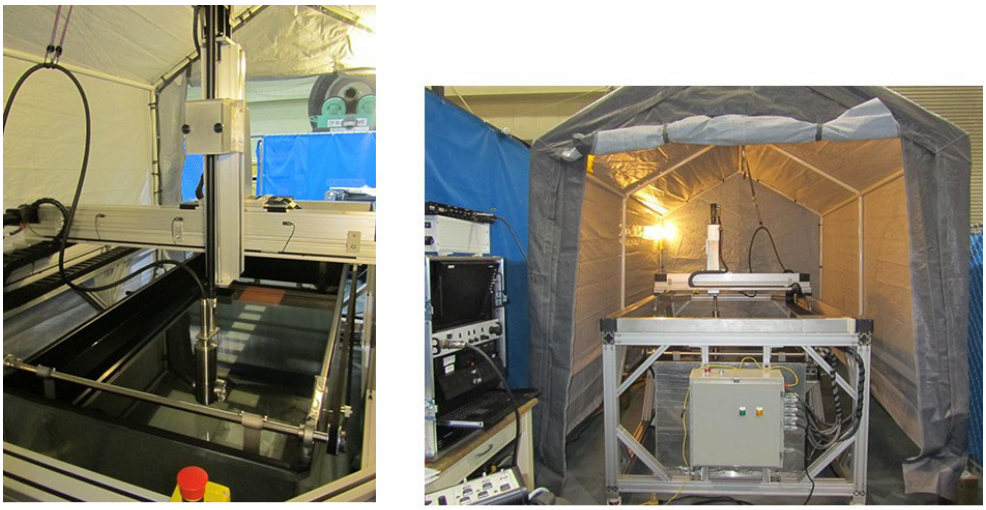
A.4 Examinations

A.4.1 General

The RV-RRT will use blind tests, where defect detection and discrimination shall be demonstrated.

A.4.2 Available Equipment

All tests will be conducted with the specimens located underwater in a tank, with a blackout tent used to exclude external (or ambient) lighting to the extent possible (Figure A-2). The camera will be mounted on an X-Y-Z-Θ scanner controlled by a joystick. In addition, the scanner has the ability to tilt the camera in a controlled fashion. The scanner is designed to hold a 25.4 mm to 43.2 mm (1 in. to 1.7 in.) camera handling pole. This scanner allows for precise control.



(a) (b)

(c)

Figure A-2 (a) Scanner shown with camera. (b) Black-out tent. (c) Diagram showing top view of scanner and test specimen layout.

A.4.3 Sample Setup for Scanning

Using the tanks and scanner, the following sample setup will be used for the RV-RRT. Samples would be placed side-by-side, on a platform in the tank. This arrangement will allow for the scanner to examine each sample sequentially in the X-direction (and scan from one end of the sample to the other in the Y-direction).

A.4.4 Monitoring

For the blind inspection it is very important to ensure the security of test specimens and examination data. The RV-RRT requires that test administrators are present during data acquisition to ensure the only information available to teams about flaws is that which they have acquired by following their examination procedure, that all steps are followed in the order specified in the procedure, and that no deviations occur. All of the inspection team members are also required to sign an Agreement to Maintain Data Confidentiality" (see Attachment A2) and include these with the datasheets.

Results from blind inspections shall be noted by the test-takers on the provided RV-RRT Data Sheets (Attachment A3).

At the end of the inspection, all of the data from storage media provided by a team will be transferred to the test administrator to maintain security of the test.

A.4.4.1 Deviations

If deviations from the examination procedure occur, they need to be documented.

If deviations from the procedure are necessary to fulfil the requirements, the following steps must be taken:

- Carefully address all deviations to the procedure.
- Date and document all changes that have to be addressed together with a statement about the reason why the deviation is necessary.
- If the procedure does not describe all of the steps in detail, then the test administrator must note this.

Note: Detailed documentation regarding deviations is necessary to help ensure that test administrators can receive counsel from other members on the RV-RRT team if necessary, and the right decision can be made to resolve issues.

Note: This information will also be provided to all other test administrators so that if similar problems are encountered, these can be handled in a consistent manner. A designated person (test administrator) always has the responsibility for final decision in the case of any dispute.

A.4.5 Overview of RV-RRT Process

This section provides an overview of the envisioned blind test process for RV-RRT.

A.4.5.1 Pre-examination Setup

The inspection team arrives on-site, and installs the camera and equipment in the secure area. The tank is filled with water and the inspector verifies proper functioning of cameras and associated equipment using non-round-robin specimens and/or resolution standards. The inspector will also be required to provide, sufficiently in advance of the test, a copy of the standard examination procedure (EP) that he/she will use (see Section A.3). Limitations on camera angle, lighting, distance, and scan speed as required by the examination procedure will be discussed and demonstrated by the inspector using the manipulator. Both camera and platform are underwater. The inspection team will also be given some time to become familiar with the manipulator controls. If the inspector performs a resolution standard check (calibration check) as required by his/her EP, the data from this check should be recorded.

A.4.5.2 Blind Round Robin Test

The following process is envisioned for the blind test.

1. Test administrator provides pre-test briefing to inspector. Briefing will consist of overview of test protocol as well as any limits imposed on the inspector (such as time allowed for inspecting a single test specimen).
2. Test administrator clears area, then loads one or more test specimens into tank. All ambient lighting inside tent is turned off, and the tent flap closed.
3. Inspector begins examination by positioning camera over first test specimen.
4. Test administrator notes start time for test specimen.
5. Inspector uses scanner controls to scan camera over the desired inspection region (weld crown and adjacent base material) on test specimen. Typically, inspector is expected to examine (screen) the weld from both sides (upstream and downstream), identify any indications, and evaluate the indications to determine whether they constitute cracks or non-relevant surface features.
6. As a part of the detection and discrimination process, the inspector may examine indications from different angles and adjust lighting on the camera as permitted by the examination procedure. If camera has pan-tilt-zoom capability, inspector may elect to use these as well, as permitted by the examination procedure.
7. All examination data will be recorded (video and audio) by the inspector. However, inspector is not allowed to review video data for detection or discrimination of flaws. All flaw calls will be made using the live video feed.
8. Inspector records any indications he/she considers cracks onto a standard data sheet (to be provided by test administrator). Locations of cracks will be determined using the markers/rulers on the specimens. The data sheet will also have space for inspector to record additional comments (pan-tilt-zoom used or not, lighting used, etc.). In addition, inspector will record still images of any indication that is called a crack.
9. Inspector turns in data sheet to test administrator. Test administrator records stop time for test specimen. Test administrator verifies that the data sheet for the test specimen is complete, and the required information has been filled in. The test administrator shall not evaluate the accuracy of the indications recorded on the data sheet (using live or recorded data) at this stage. However, he/she should review the data sheet for completeness.
10. Inspector begins evaluating next test specimen. He/she repeats steps 3–9 for each test specimen in the tank.
11. When inspector completes all test specimens currently loaded in tank, test administrator clears area and replaces the test specimens in tank with new ones (assuming there are more test specimens remaining in the test sequence). The specimens that were in the tank are placed in secure storage.
12. Inspector repeats examination process (steps 2–10). When complete, test administrator may replace test specimens with new ones (step 11). The process is continued until all test specimens in this test sequence are completed.

13. Once all tests are complete, all test specimens are secured. Inspector then performs final resolution standard check (exit calibration check) as required by the examination procedure. Inspector is then allowed to enter blackout tent, and remove camera equipment.
14. Steps 1–13 are repeated for each inspector/team.

A.4.6 Rules for Conducting the RV-RRT

During the performance of the RV-RRT, the following items shall be included and examined by the test administrator.

- A single test is defined as the inspection of all specimens provided to the inspector in some predetermined sequence. A single test will comprise approximately 45 test specimens.
- Detection and discrimination shall be demonstrated and documented.
- Personnel that are performing the examination shall be qualified in accordance with the requirements stated in the EP.
- A single inspector will be allowed to take each test. The inspector will be responsible for review of the camera feed and documentation of any calls. A second person (operator) will be responsible for camera positioning by controlling the manipulator. The operator will only be allowed to operate the manipulator, and will not be allowed to communicate with the inspector beyond the minimum necessary to ensure proper positioning of the camera. Communications will be monitored to ensure that under no circumstance can the inspector ask for advice or guidance from another member of the team or from the operator.
- The test administrator can, if necessary, provide initial guidance relative to positioning the camera on the first specimen (to ensure that the camera is properly positioned prior to beginning the test).
- The equipment specified in the procedure shall correspond with the equipment used during the inspection.
- The inspector shall follow all steps included in the procedure and any instructions or manuals.
- All tests will be conducted in a tank, with a blackout tent used to exclude external (or ambient) lighting. The camera will be mounted on a multi-axis scanner controlled by a joystick. The camera will be scanned over each specimen to inspect the designated region of interest.
- The inspector will be given 20 minutes to inspect each specimen.
- The inspector will have the freedom to adjust lighting and the angle and magnification of the camera to the extent defined by the examination procedure. If the camera has pan, tilt, or zoom capabilities, these may be used to the extent defined by the examination procedure. Supplemental lighting shall not be permitted.
- The inspector shall not change cameras during a single test. Separate tests using a different camera may be performed if previously cleared with the test administrator and if additional time is available.
- The inspector will not be allowed to perform an overview inspection of a specimen prior to performing the documented (i.e., recorded) examination.

- If the examination procedure requires a resolution check, the inspector will record all data related to the resolution check. These data will be recorded as video (and if appropriate, still images). In addition, the inspector will fill out the Resolution Check Data Sheet (Attachment A4). The time to perform resolution checks does not count as inspection time.
- All inspection results will be reported using the standard RV-RRT data reporting form (Attachment A3). Copies of data reporting forms will be provided by the test administrator to the inspector prior to beginning each test.
- In addition to the inspection results, any other reporting as required by the examination procedure should also be performed, explained, and presented to the test administrators by the inspection team after completion of test plate examination.
- The inspector will record video documenting the entire examination process for each specimen. The recorded video shall also include footage of the specimen label prior to starting an examination of a test specimen. Prior to starting an examination of a test specimen, the inspector shall also take a still picture of the specimen label and save this image.
- Audio commentary should also be recorded to help document the detection and identification process.
- Data shall be stored on agreed-upon storage media in agreed-upon formats. This will include images and video, as well as any other form of data (written documents, etc.).
- All flaw calls will be made using only the live feed from the camera. The inspector will NOT be allowed to make or change calls by reviewing the recorded data (video or still pictures).
- If a call is made, the inspector will make a still image of the flaw and identify the region containing the flaws using the measuring scale on the specimen.
- All inspectors, test administrators, and other authorized personnel shall adhere to the round-robin security protocol.
- Each inspector will be provided with 1–2 specimens that can be used for training purposes. These specimens will be provided before the first test set is provided.

A.4.7 Location for Remote Visual Round Robin Testing

Inspections will be conducted at the EPRI NDE Center, Charlotte, North Carolina.

Note: Test administrators need access to an area for blind inspections, which can be locked, as well as office space for their exclusive access.

A.4.8 Start-up for RV-RRT

A letter shall be sent by test administrators to participating inspection teams to explain the RV-RRT and how it will be performed. This shall be done sufficiently in advance of the test.

A.4.9 Coordinate Systems and Reporting Units

To ensure that testing teams report all data for detected defects in test pieces in a uniform way, a standard coordinate system (Attachment A5) will be adopted. A measuring scale shall be etched, taped, or otherwise mounted on the test specimens to assist the inspector in recording the location of any detected cracks.

A.4.10 Evaluation of RV-RRT Results

Evaluation of reported results should be performed after each participant has completed the RV-RRT. This is to ensure that the evaluation of results from all participants is completed in a timely manner.

Evaluation criteria for RV-RRT are being developed jointly by EPRI and PNNL. This information shall be included in this document when it is available (no later than the start of the RV-RRT).

A.5 Reporting

Test administrators have to fill in a checklist (see Attachment A6) and ensure that everything during the blind test has been done systematically and in accordance with the inspection procedure.

Standard examination data reports developed for recording data from the inspection team shall be used for entering into a database for subsequent analysis. These forms, shown in Attachment A3 and Attachment A6, shall be signed and dated by the inspection team and the test administrator, respectively. The test administrator is to ensure that all data have been entered into the form. The test administrator is responsible for these signed forms and their transfer to the evaluation team.

ATTACHMENTS TO APPENDIX A

- A1 – Examination Procedure Summary
- A2 – Agreement to Keep RV-RRT Test Information Confidential
- A3 – RV-RRT Inspection Data Sheets
- A4 – Resolution Check Data Sheet
- A5 – Coordinates
- A6 – Checklist for Test Administrators

Examination Procedure Summary Attachment A1

Team Code:

Inspection Procedure Summary:

(This will be developed by the test administrator and then reviewed by the inspection team to get their approval that the TA is accurately representing the procedure.)

Agreement to Keep RV-RRT Test Information Confidential

Attachment A2

By my signature below, I confirm that I have read and understand Section A.2.7 (Data Security) of the RV-RRT protocol, and agree to comply with all aspects of the data security requirements.

TEAM SIGN:

Team Member Name (Printed)	**Team Member Name (Signature)**	**Date**
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

ROUND ROBIN DATA SHEET - PHASE II

Vendor ID: _____

TEST INFORMATION

VT Inspector: _____

Manipulator Operator: _____

Procedure Number: _____

Date: _____

SYSTEM INFORMATION

Camera ID: _____ ☐ Color ☐ B&W

Zoom Capable: ☐ Yes ☐ No

Camera Resolution: _____

Lighting Type (on camera): ☐ Halogen ☐ LED

Lighting Comments: _____

Data Acquisition System: _____

Monitor Manufacturer: _____ Model: _____ Resolution: _____ Size: _____

SRCS Serial Number: _____

SRCS Description: _____

DATA STORAGE:

Recording media used: ☐ Hard drive ☐ DVD ☐ Other

Media Identification Label: _____

COMMENTS:

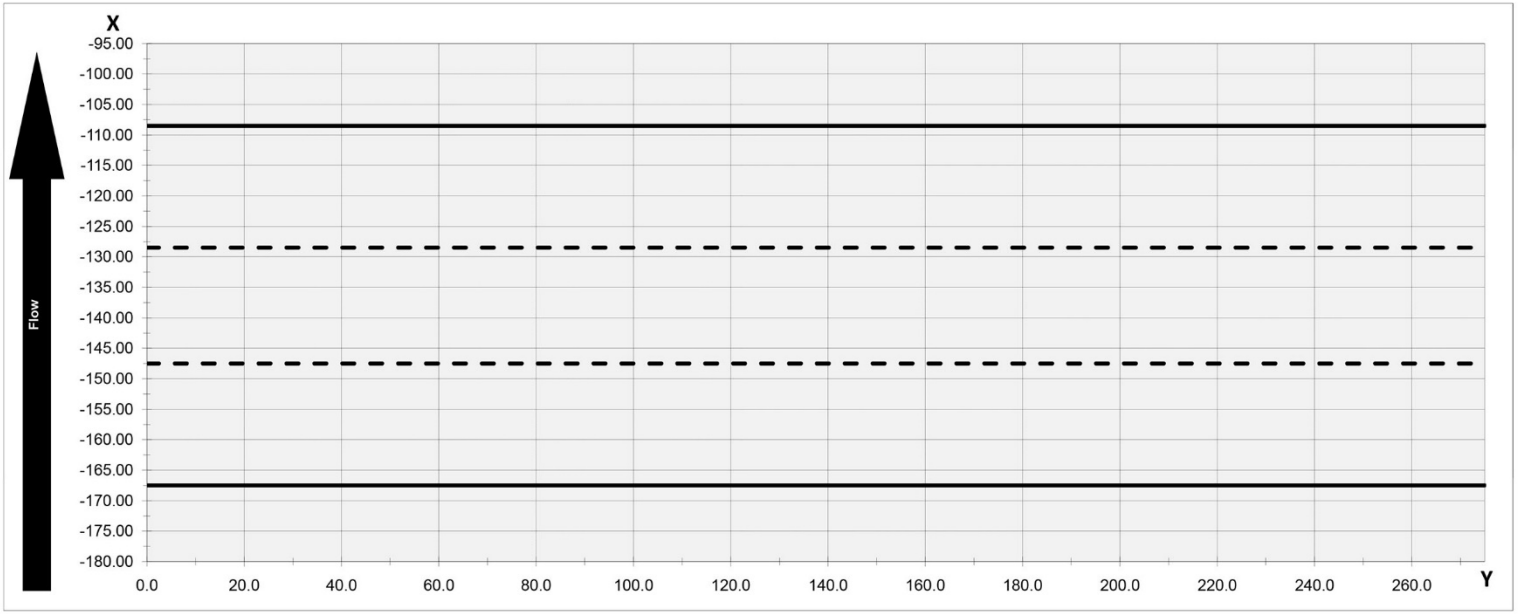
SIGNATURE: _____ **Date:** _____

Candidate Indication Data Report

Sample ID: EXAMPLE **Units: mm** **Candidate:** _____ **Date:** _____

Examination Time

Exam Start: _____ **Exam End:** _____



	Candidate Flaw Information												
Y start													
Y stop													
X start													
X stop													
Length													

Sample Information	
Weld Length	275.0
Weld Center Line	-138.0
Weld Width	19.0

Candidate Notes: _____

_____

_____

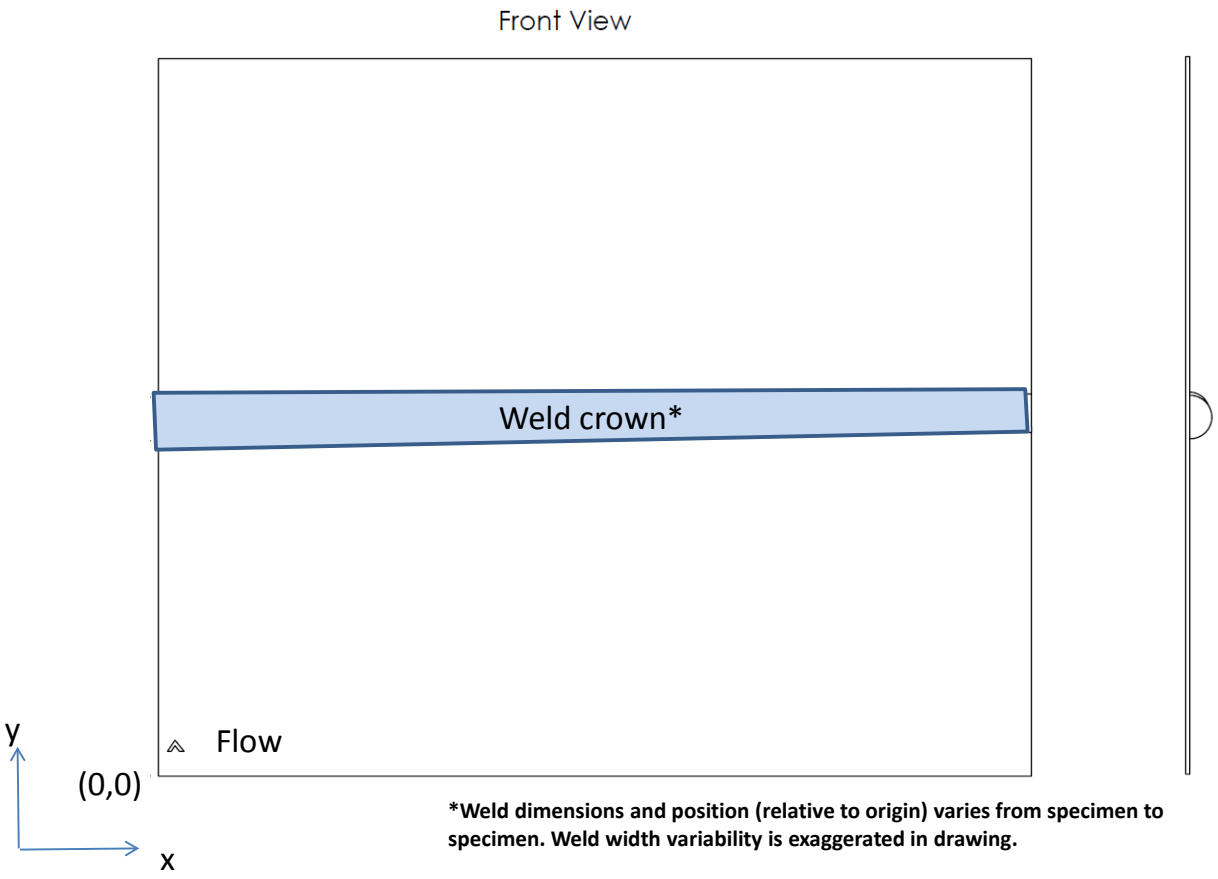
Candidate Signature: _____

Resolution Check Data Sheet — Attachment A4

SRCS Resolution Data Log				
Vendor:				
	Time	**Date**	**Len-to-Target Distance**	**Comments**
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
SIGNATURE: _____			**Date:** _____	

Coordinates Attachment A5

Stainless Steel Specimens



***Weld dimensions and position (relative to origin) varies from specimen to specimen. Weld width variability is exaggerated in drawing.**

Ceramic Specimens

Checklist for Test Administrators Attachment A6

The purpose of this checklist is to facilitate a systematic process for conducting the RV-RRT. This checklist presupposes that data acquisition and data review are performed at the same time (i.e., using live camera feed).

Item	Test Administrator Initials	Date	Comments
List of team members			
Contact person for team			
Inspection Procedure – Technical description submitted for review			
Inspection Procedure Summary document completed (Att. A1 in PROTOCOL)			
Test administrator review of test specimens and order of test specimens			
Test administrator review setup of equipment and compares to procedure			
Test administrator observes calibration of equipment and compares to procedure			
Test administrator review data acquisition and review, and compares to procedure			
Test administrator reviews inspector fill in of Data Sheets			
If deviations from procedure are found, make a clear comment and justify deviation, as well as assuring that it is dated and fully documented.			
Send all data sheets to PNNL/EPRI for analysis			

APPENDIX B PHASE III PROTOCOL

B.1 Summary

The United States Nuclear Regulatory Commission (NRC), in cooperation with the Electric Power Research Institute (EPRI), is conducting an evaluation of the reliability of remote visual testing (EVT-1) currently used for in-vessel visual inspection (IVVI). A limited round-robin test was conducted during Phase I of this evaluation. The results of the Phase I test, along with a subsequent parametric study, were used to design a more extensive Phase II round-robin test (designated as the RV-RRT in this document). Phase II results pointed to the possibility of improving procedures for IVVI to enhance the ability to discriminate between cracks and non-relevant indications such as surface features. These results, along with a subsequent parametric study, were used to design a Phase III round-robin test (designated as the RV-RRT-3 [Remote Visual Round-Robin Test Phase III] in this document).

The project will focus on examinations in the welded region of large components (using flat plates with simulated welds). The study will assess crack detection and discrimination, where discrimination means distinguishing between cracks and non-relevant surface features.

The project time schedule for completion of this set of round-robin testing is first quarter of 2016.

This document describes the scope, prerequisites, organizations, and rules for the RV-RRT-3 during Phase III of the research.

B.2 General

B.2.1 Scope of Remote Visual (RV) Round-Robin Test (RRT) Phase III

The purpose of this PROTOCOL document is to describe the scope, prerequisites, organizations, and rules for the Remote Visual Round-Robin Test Phase III (RV-RRT-3).

The NRC, in cooperation with the Electric Power Research Institute (EPRI), is conducting an evaluation of the reliability of remote visual testing (EVT-1) currently used for in-vessel visual inspection (IVVI). A limited round-robin test was conducted during Phase I of this evaluation. The results of the Phase I test, along with a subsequent parametric study, were used to design a more extensive Phase II round-robin test (designated as the RV-RRT in this document). Phase II results pointed to the possibility of improving procedures for IVVI to enhance the ability to discriminate between cracks and non-relevant indications such as surface features. These results, along with a subsequent parametric study, were used to design a Phase III round-robin test (designated as the RV-RRT-3 in this document).

The RV-RRT-3 comprises a study of the effectiveness of remote visual non-destructive examination (NDE) techniques currently used for IVVI when combined with proposed procedure improvements. The round-robin testing is intended to provide an assessment of commercially applied examination procedures using a blind testing methodology. The goal is to assess the performance of improved procedures with qualified personnel. The results from the RV-RRT-3 will be used to identify areas for future improvement, if needed.

The project will focus on remote visual examinations in welded regions, using flat plates with weld crowns (as-welded or flush). The study will assess crack detection and discrimination, where discrimination means distinguishing between cracks and non-relevant surface features.

The RV-RRT-3 will be carried out as a coordinated effort between the Pacific Northwest National Laboratory (PNNL) and EPRI, with participating examination teams from the United States with direct field experience with IVVI in nuclear facilities.

B.2.2 RV-RRT-3 Objectives

The goal of the Remote Visual Round-Robin Test Phase III is to assess the performance of commercially applied examination procedures augmented with improvements for enhancing detectability of cracks with qualified personnel and identify areas for future improvement, if needed. Current examination procedures typically permit an independent review of recorded examination data including reported indications and a re-examination of questionable indications/areas.

The RV-RRT-3 has the following objectives:

- Identify and quantitatively assess enhancements to remote visual examination techniques for detecting and characterizing flaws in test specimens.
  - Evaluate improvements to commercially applied examination procedures for their effectiveness.
    - Quantify (in terms of improvement to flaw detection rates) the impact of secondary review of all recorded examination data
    - Quantify the level of image degradation (if any) in recorded data.
  - Quantify procedure performance in terms of probability of detection (POD) and determine the effect that certain important factors have on POD. Important factors for Phase III include:
    - Crack opening displacement/dimension (COD)
    - Crack length
    - Crack detection in the presence of surface irregularities or blemishes.

B.2.2.1 Expected Outcomes

The RV-RRT-3 data are expected to provide a better overall understanding of the performance of augmented commercially applied remote visual examination procedures and the critical factors that affect crack detection performance.

In addition, from the RV-RRT-3 data, be able to calculate the following:

- POD curves for each participating examination team as a function of flaw size (COD and length).
- If the results from more than one team can be grouped (for instance, if the examination procedures used allow such grouping), POD curves will be created for these groups.

- Identification of significant differences in POD related to important variables:
  - Examination procedure, flaw type, and orientation.
  - Evaluation of false call probability (FCP).

B.2.2.2 Limitations

The test specimens used in this round robin (and described in Section B.2.6) represent welds similar to that found in some reactor internal components. The specimens represent a specific color (patina) in reactor internals (natural stainless steel) and may not be representative of the diverse color variations of internals surfaces. The effect of other configurations and surface conditions on test results is beyond the scope of this round robin. As typically permitted in current examination procedures, an independent review of recorded examination data including reported indications may be performed as part of the test and a re-examination conducted of questionable indications/areas. However, some limitations may be imposed related to number of re-looks, etc. Finally, the RV-RRT-3 is not designed to assess certain variables that may impact detection performance in remote visual testing. These are:

- Oxide build-up on internal components
- Thermal distortion
- Water currents and clarity
- Radiation effects on camera video quality
- Limited accessibility (component configurations and camera size)
- Camera delivery systems
- Personnel qualification levels
- The angle of view limits for Code VT-1 examinations
- Application to geometrical configurations other than flat.

B.2.3 RV-RRT-3 Project Organization

NRC Program Manager:
Ms. Carol Nove
U.S. Nuclear Regulatory Commission
RES
Washington, DC 20555-0001
USA
Phone: 301-251-7664
Email: carol.nove@nrc.gov

Industry Steering Committee Chair:
Mr. Michael Oliveri
PSEG Nuclear, LLC
End of Alloway Creek Neck Rd
Hancocks Bridge, NJ 08038
USA
Phone: 856-339-3538
Email: michael.oliveri@pseg.com

RV-RRT-3 Project Contact Persons:

Dr. Pradeep Ramuhalli, PNNL
Scientist/Engineer
P.O. Box 999, MSIN K5/26
Richland, WA 99352
USA
Phone: 509-375-2763
Email: pradeep.ramuhalli@pnnl.gov

Mr. Jeff Landrum, EPRI
1300 West WT Harris Blvd
Charlotte, NC 28262
Phone: 704-595-2553
Email: jlandrum@epri.com

Mr. Michael T. Anderson
Scientist/Engineer
P.O. Box 999, MSIN K5/26
Richland, WA 99352
USA
Phone: 509-375-2523
Email: michael.anderson@pnnl.gov

Mr. John Lindberg, EPRI
Program Manager – NDE Innovation
1300 West WT Harris Blvd.
Charlotte, NC 28262
Phone: 704-595-2625
Email: jlindberg@epri.com

Industry Steering Committee:

- Michael Oliveri, Chair Ad-Hoc Committee – APC 4, Public Service Electric and Gas
- Kevin Hacker – NDE IC Chair, Dominion
- Tim Wells – MRP IIG Chair, Southern Nuclear
- Marc Brooks – NDE Reliability TAC, Detroit Edison
- Harry L. Smith – APC 4, Exelon
- Chris McKean- BWRVIP IFG Technical Chair, Exelon
- Dan Nowakowski – MRP ITG, NextEra

Test Administrators:

- PNNL: Dr. Pradeep Ramuhalli
- PNNL: Mr. Michael Anderson
- PNNL: Ms. Susan Crawford
- EPRI: Mr. Jeff Landrum
- EPRI: Mr. Chris Joffe
- EPRI: Mr. John Lindberg
- EPRI Contractor: Mr. Jonathan Buttram

B.2.4 Examination Teams

The following examination companies are participating in the remote visual round-robin test.

Company
AREVA
Westinghouse
WesDyne
General Electric Hitachi
IHI Southwest

B.2.5 Examination Personnel

Each examination team will consist of two separate analysts. An additional individual (henceforth referred to as the "Operator") may be used for camera manipulation. Each analyst shall be qualified in accordance with the requirements set forth in their Examination Procedures (see Section B.3.2), and shall be familiar with the requirements of ASME Code VT-1 examination and EVT-1, as well as their Examination Procedure. One analyst (henceforth referred to as the "Primary Analyst") will perform the primary evaluation (see Section B.4.6 for a description of the process, and Section B.5 for the associated Rules). The second analyst ("Secondary Analyst") will use the recorded data and the data sheets from the primary analyst to review and confirm the results of the examination.

B.2.6 RV-RRT-3 Specimen Information

B.2.6.1 Test Objects

Category	Typical Weldment	Comments/Notes
Stainless Steel Flat Plates	Butt weld	Welds crowns may be narrow or wide. Some weld crowns may be ground flush. Specimens have surface features such as grinding and tooling marks (scratches).

The RV-RRT-3 will consist of a sufficient number of cracked and blank grading units to enable calculation of POD and associated confidence bounds.

The specimens will be rectangular plates (dimension varies) of stainless steel with either a narrow or wide weld crown. Some weld crowns may be removed via grinding. The specimens will be clean, but not shiny, for the examinations. No additional cleaning of the specimens will be allowed. Bubbles that may form on test specimens may be removed by gentle brushing if requested.

A measuring scale will be etched, taped, or otherwise mounted on the specimen to assist the inspector in recording the location of any detected cracks.

Each sample will be labelled using a unique identifier (ALIAS name) that will be visible to the inspector.

Note: Unless there are mitigating circumstances, as approved by the test administrator prior to taking the test, participating teams must examine all test specimens provided.

B.2.6.2 Examination Region

The region of interest will be the base materials adjacent to the simulated weld on both the upstream and downstream sides, as well as the weld itself. The examination region will be identified for each specimen by the area between the measuring tapes located on both sides of the welded areas (examination area typically includes weld width plus adjacent one inch of base material on either side).

B.2.6.3 Defect Specification

The expected degradation mechanisms in internals components are stress corrosion cracking and fatigue cracking. The test specimens may contain cracking anywhere within the examination region with defect orientation being transverse or longitudinal to the welding direction.

Discrimination between cracking defects and non-relevant surface features (such as scratch marks) shall be part of the evaluation process. Using the measuring scales provided on the specimen, flaw lengths shall be documented onto the data sheet. In the case of flaws oriented perpendicular to the weld, location along the weld axis is documented using the visible scales—flaw length is estimated based on visible landmarks and shown on the data sheet.

B.2.7 Data Security

All information concerning the blind tests is considered to be confidential and shall therefore be dealt with as such. Specifically, all parties participating in the RV-RRT-3, and/or in evaluating the results from the RV-RRT-3, shall not release or discuss any data, results, papers or data media, or any other information, to anyone not authorized for that type of information without prior approval from NRC-RES and the Industry Ad-Hoc Committee.

Authorized personnel are those who participate in the RV-RRT-3 and the subsequent data analysis (NRC, PNNL, EPRI, Ad-hoc committee members, and the test administrators).

Summary results from the RV-RRT-3 will be published after the analysis of the results is complete. The identity of the examination teams shall be anonymous throughout the testing period and afterward in published reports and documents.

All personnel in the RV-RRT-3 project team, test administrators, and examination teams conducting the RV-RRT-3 must comply with this protocol.

The following restrictions shall be applied:

- All papers and information, including scrap papers and data media, must be handed over to test administrators and those who are responsible for the RV-RRT-3 activities. No unauthorized person may remove such information from the test facilities.
- During the RV-RRT-3, no Internet connections, electronic devices, or wireless devices (cell phones, iPhones, etc.) are allowed.

- Copying of data or data transformation is not allowed for any purpose without prior permission from authorized personnel (test administrators). The only exception is the copying of data for the purposes of secondary review.
- Procedures for making a copy of the data for the purpose of secondary review need to be approved by the test administrators. Media used for recording or copying data for secondary review shall be supplied by the test administrator.
- Removable data storage media (such as memory sticks) are not allowed in the test facilities, except for those accepted by test administrators to be used during the examination.
- The RV-RRT-3 data shall not be recorded onto external media (such as computer disk drives) brought on-site by the examination team. The test administrator shall supply storage media for recording the RV-RRT-3 data.
- Neither the test administrators nor the participants in the RV-RRT-3 may discuss flaws or results with other participants in the RV-RRT-3.
- A computer may be provided by the test administrators for the Secondary Analyst to watch the recorded videos. If desired, a team may supply a computer; however, because it will be located in a secure area, it will fall under Performance Demonstration Initiative (PDI) rules for security and the hard drive will be wiped cleaned of ALL data upon exit of the secure area.

Test specimens, test results, and teams will be assigned ALIAS names for recording data and results. The ALIAS names are produced by PNNL and EPRI, and are designed to conceal the identity of the participants and the items listed.

Only authorized personnel will have the knowledge of the real team names, test results, or test specimens used.

- All data generated in the RV-RRT-3 will be provided to the test administrators.
- All test administrators must have a proven independence and impartiality status with respect to the participating teams.
- All test administrators need to have an approved back up in case they become sick or job conflicts prevent them from being able to perform their duties.

B.3 Document Review

B.3.1 Introduction

Each test administrator will review the Examination Procedures (EP) developed by the participating examination teams for the RV-RRT-3. This requirement is put in place to facilitate review and analysis of the results of RV-RRT-3, and to quickly get answers to questions that arise during the testing, data analysis, and review process.

Examination Procedures should be sent to the test administrator sufficiently in advance of the scheduled test period. Test administrators will review these documents and write a summary document called Procedure Summary (see Attachment B1), for review by PNNL and EPRI. This document will summarize the techniques for detection and flaw discrimination. The purpose of this document is to facilitate review and assessment of data (during data analysis) without having to

read all examination procedures in detail. The Procedure Summary shall not contain any proprietary information in the examination procedures.

B.3.2 Examination Procedure

Examination Procedures should contain all relevant information regarding the preparation, performance, and reporting of remote visual examination. The examination procedure should describe **what to do** and **how to do it**. It should be clearly stated in the scope of the examination procedure within which limits it is valid (for example, material type, surface finish, etc.).

The purpose of the RV-RRT-3 is to assess enhancements to commercial procedures for the purpose of increasing the ability to discriminate between cracks and non-relevant surface features. However, further adaptations could be necessary to accommodate/fit the test specimen geometries and the test setup.

B.3.3 Augmented Examination Procedure

The test administrators will provide supplementary guidance (see Attachment B2) in the form of a set of recommended best practices for remote visual examination to each of the participating examination teams. This list of best practices will constitute the enhancements to the EP used by the participating teams. The resulting EP with the supplementary guidance will be referred to in the rest of this document as the Augmented EP.

Note: With the exception of providing the supplementary guidance, it is not the role of test administrators to make any comment about the procedures used by the participants.

B.4 Examinations

B.4.1 General

The RV-RRT-3 will use blind tests, where defect detection and discrimination shall be demonstrated.

B.4.2 Location for Remote Visual Round-Robin Testing

Examinations will be conducted at the EPRI NDE Center, Charlotte, North Carolina.

Note: Test administrators need access to an area for blind examinations, which can be locked, as well as office space for their exclusive access.

B.4.3 Start-up for RV-RRT-3

The examination protocol shall be provided to participating examination teams to explain the RV-RRT-3 and how it will be performed. The protocol should be provided sufficiently in advance of actual testing.

B.4.4 Available Equipment

All tests will be conducted with the specimens located underwater in a tank, with a blackout tent used to exclude ambient lighting to the extent possible (Figure B-1). The camera will be mounted on an X-Y-Z-Θ scanner controlled by a joystick. In addition, the scanner has the ability to tilt the

camera in a controlled fashion. The scanner is designed to hold a 25.4 mm to 43.2 mm (1 in. to 1.7 in.) camera handling pole. This scanner allows for precise control of examination cameras.

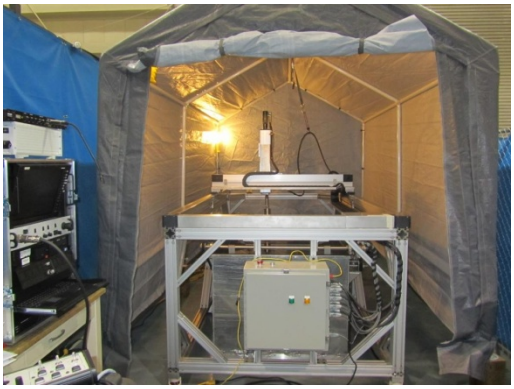


Figure B-1 Black-out Tent with Tank and Scanner

B.4.5 Sample Setup for Scanning

Using the tanks and scanner, the following sample setup will be used for the RV-RRT-3. Samples would be placed in an individual specimen holder mounted to the side of the tank. This arrangement will allow for the scanner to examine each sample sequentially in the X-direction (and scan from one end of the sample to the other in the Y-direction).

B.4.6 Monitoring

For the blind examination it is very important to ensure the security of test specimens and examination data. The RV-RRT-3 requires that test administrators are present during data acquisition to ensure the only information available to teams about flaws is that which they have acquired by following their examination procedure, that all steps are followed in the order specified in the procedure, and that no deviations occur. All of the examination team members are also required to sign an Agreement to Maintain Data Confidentiality" (see Attachment B3) and include these with the datasheets.

Results from blind examinations shall be noted by the test-takers on the provided RV-RRT-3 Data Sheets (Attachment B4).

At the end of the examination, all of the data from storage media provided by a team will be transferred to the test administrator to maintain security of the test.

B.4.6.1 Deviations

If deviations from the supplied Augmented EP occur, they need to be documented.

If deviations from the procedure are necessary to fulfil the requirements, the following steps must be taken:

- Carefully address all deviations to the procedure.
- Date and document all changes that have to be addressed together with a statement about the reason why the deviation is necessary.

- If the procedure does not describe all of the steps in detail, then the test administrator must note this.

Note: Detailed documentation regarding deviations is necessary to help ensure that test administrators can receive counsel from other members on the RV-RRT-3 team if necessary, and the right decision can be made to resolve issues.

Note: This information will also be provided to all other test administrators so that if similar problems are encountered, these can be handled in a consistent manner. A designated person (test administrator) always has the responsibility for final decision in the case of any dispute.

B.4.7 Overview of RV-RRT-3 Process

This section provides an overview of the envisioned blind test process for RV-RRT-3.

B.4.7.1 Pre-examination Setup

The examination team arrives on-site, and installs the camera and equipment in the secure area. The tank is filled with water and the inspector verifies proper functioning of cameras and associated equipment using non-round robin specimens and/or resolution standards. The inspector will also be required to provide, sufficiently in advance of the test, a copy of the standard EP that he/she will use (see Section B.3.2). Limitations on camera angle, lighting, distance, and scan speed as required by the examination procedure will be discussed and demonstrated by the inspector using the manipulator. Both camera and specimen holder are underwater. The examination team will also be given some time to become familiar with the manipulator controls. If the inspector performs a resolution standard check (calibration check) as required by his/her EP, the data from this check should be recorded.

In addition to any EP-required resolution standards, the inspector will also record data from a resolution standard provided by the test administrator. Data from this auxiliary resolution standard will be used to determine if there is any degradation of image quality in recorded data. A standard procedure (Attachment B5) and data sheet (Attachment B6) for recording data from the auxiliary resolution standard will be provided by the test administrator.

B.4.7.2 Practice Specimens

Prior to beginning the blind round-robin test, the examination team will be provided with the supplementary guidance by the test administrator. Sufficient time will be provided to the examination team to familiarize itself with the supplementary guidance. In some cases, this guidance may be sent prior to the examination team's arrival on site to take the test. A set of practice specimens will be provided to the examination team prior to the round-robin test specimens to facilitate practice application of the supplementary guidance.

B.4.7.3 Blind Round Robin Test

The following process is envisioned for the blind test, once the pre-examination setup is complete.

1. Test administrator provides pre-test briefing to examination team. Briefing will consist of overview of test protocol as well as any limits imposed on the team (such as time allowed for examination and primary evaluation of a single test specimen).

- **Objective**: Ensure examination team is familiar with enhancements to EP and recommended best practices.

2. Test administrator clears area, then loads one or more test specimens into tank. All ambient lighting inside tent is turned off, and the tent flap closed.
   - **Objective**: Maintain security of test specimens.
3. Primary Analyst, with the Operator's assistance, begins examination by positioning camera in a position to view the first test specimen.
   - **Objective**: Begin primary examination and evaluation of test specimen and document results on supplied data sheets.
   - **Note**: Steps 4–9 are part of the primary evaluation process.
4. Test administrator notes start time for test specimen.
   - **Objective**: In combination with Step 9, document time for inspecting a single specimen.
5. Camera operator uses scanner controls to scan camera viewing the desired examination region (weld crown and adjacent base material) on test specimen. Typically, Primary Analyst is expected to examine the weld and both adjacent sides (upstream and downstream), identify any indications, and evaluate the indications to determine whether they constitute cracks or non-relevant surface features.
6. As a part of the detection and discrimination process, the Primary Analyst may examine indications from different angles while adjusting lighting on the camera as permitted by the examination procedure. Auxiliary lighting, as permitted by the Augmented EP, may be used if necessary as part of the evaluation process. If camera has pan-tilt-zoom capability, Primary Analyst may elect to use these as well, as permitted by the Augmented EP.
7. All examination data will be recorded (video and audio) by the Primary Analyst. However, Primary Analyst is not allowed to review video data for detection or discrimination of flaws at this stage (primary evaluation). All flaw calls by Primary Analyst will be made using the live video feed.
   - **Objective**: To make available recorded data for secondary review (Step 10).
8. Primary Analyst records any indications he/she considers cracks onto a standard data sheet (to be provided by test administrator). Locations of cracks will be determined using the markers/rulers on the specimens. The data sheet will also have space for Primary Analyst to record additional comments (pan-tilt-zoom used or not, lighting used, etc.). In addition, inspector will record still images of any indication that is called a crack.
   - **Objective**: To document results and variables used during the examination process, for use in the eventual analysis of RV-RRT-3 data.
9. Primary Analyst turns in data sheet to test administrator. Test administrator records stop time for test specimen. Test administrator verifies that the data sheet for the test specimen is complete, and the required information has been filled in. The test administrator may not evaluate the accuracy of the indications recorded on the data sheet (using live or recorded data) at this stage. However, he/she should review the data sheet for completeness.
10. When Primary Analyst completes test specimen currently loaded in tank, test administrator clears area and replaces the test specimen in tank with a new one (assuming there are more test specimens remaining in the test sequence). The specimen that was in the tank is placed in secure storage.

11. Primary Analyst begins evaluating next test specimen. He/she repeats steps 3–9 for each test specimen. When complete, test administrator may replace test specimens with new ones (step 10). The process is continued until all test specimens in this test sequence are completed.
12. Secondary review: The data sheets from step 9 for each of the test specimens and the associated recorded examination data are provided to the Secondary Analyst. A secondary review of ALL recorded data is performed solely by the Secondary Analyst, and the results of this review recorded on the provided examination data sheet (Attachment B4). The time taken to review the data for a test specimen will be recorded on the data sheet by the test administrator. The data sheet will have space for the Secondary Analyst to add additional comments, including the need for a re-examination of a questionable area or indication. The secondary review process may group data sheets from several test specimens at a time to allow for efficient use of time. Any desired re-examinations will be performed after completion of primary evaluation of all test specimens.
    - **Objective**: To obtain necessary information for assessing the impact of secondary review on the detection and discrimination performance of remote visual examination.
13. Re-examination and resolution: Through the independent secondary review process, a limited number of test specimens (constituting a maximum of 20 questionable indications or specific areas containing possible indications) will be identified for re-examination by the secondary review analyst. The test administrator will re-load these specimens. The examination follows the same process as described in steps 5–9. However, during the re-examination, the Primary and Secondary Analysts may consult during the performance of this examination using the augmented EP, and disposition the indication appropriately using the live data. Results of this re-examination will be reported on the provided examination data sheets (Attachment B4).
14. Once all tests are complete, all test specimens are secured. The examination team performs a final resolution standard check (exit calibration check) as required by the examination procedure. The team is then allowed to enter the blackout tent and remove camera equipment.
15. A de-brief may be performed for the examination team, so the test administrator and the testing team can get insights into the use of the augmented EP, the primary and secondary review process, the logic used to discriminate between cracks and non-relevant indications, and the re-examination and dispositioning process applied.
16. Steps 1–15 are repeated for each examination team.

B.4.8 Coordinate Systems and Reporting Units

To ensure that testing teams report all data for detected defects in test pieces in a uniform way, a standard coordinate system (Attachment B7) will be adopted. A measuring scale shall be etched, taped, or otherwise mounted on the test specimens to assist the inspector in recording the location of any detected cracks.

B.4.9 Evaluation of RV-RRT-3 Results

Preliminary evaluation of reported results should be performed after each participating team has completed the RV-RRT-3. This is to ensure that the evaluation of results from all participants is completed in a timely manner.

Evaluation criteria for RV-RRT-3 are being developed jointly by EPRI and PNNL. This information shall be documented in a separate document when it is available (no later than the start of the RV-RRT-3).

B.4.10 Reporting

Test administrators have to fill in a checklist (see Attachment B8) and ensure that everything during the blind test has been done systematically and in accordance with the examination procedure.

Standard examination data reports developed for recording the data from the examination team shall be used for entering into a database for subsequent analysis. The test administrator is to ensure that all data have been entered into the form. The test administrator is responsible for these forms and their transfer to the evaluation team.

B.5 Rules for Conducting the RV-RRT-3

During the performance of the RV-RRT-3, the following items shall be included and examined by the test administrator.

B.5.1 Examination Personnel Roles and Responsibilities

- A single inspector (Primary Analyst) will be allowed to perform the primary examination in each test. The Primary Analyst will be responsible for review of the camera feed and documentation of any calls during the primary examination phase only.
- A second person (Operator) will be responsible for camera positioning by controlling the manipulator. The operator will only be allowed to operate the manipulator, and will not be allowed to communicate with the inspector beyond the minimum necessary to ensure proper positioning of the camera.
- A single inspector (Secondary Analyst) will be allowed to perform the independent secondary review and any subsequent re-examinations.
- The Secondary Analyst will have the authority for making the final decision on reportable indications, based on the review and re-examinations.
- Personnel that are performing the examination shall be qualified in accordance with the requirements stated in the augmented EP. For the purpose of the Phase III Round Robin, the desired approach includes the use of a qualified Level 2 as the Primary Analyst and the use of a qualified Level 3 as the reviewer (Secondary Analyst).
- All examination team members, test administrators, and other authorized personnel shall adhere to the round-robin security protocol.
- Communications between the Primary Analyst and the Operator /Secondary Analyst will be monitored to ensure that under no circumstance can the Primary Analyst ask for advice or guidance from another member of the team or from the operator.

B.5.2 Training/Practice

- The examination team will be provided with a set of specimens (approximately 5) for practice, and to ensure that the remote visual examination equipment is functional. These specimens will be provided before the first test set is provided.

- The practice specimens may be used by the test administrators to provide guided training, on the use of the supplementary guidance that augment the examination team's EP.

B.5.3 Equipment and Specimens

- A single test is defined as the examination of all specimens provided to the examination team in some predetermined sequence. A single test will comprise approximately 45 test specimens.
- Detection and discrimination shall be demonstrated and documented.
- The equipment specified in the procedure shall correspond with the equipment used during the examination.

B.5.4 Examination Procedures

- The examination team shall follow all steps included in the procedure and any instructions or manuals.
- The test administrator can, if necessary, provide initial guidance relative to positioning the camera on the first specimen (to ensure that the camera is properly positioned prior to beginning the test).
- If the EP requires a resolution check, the analyst (primary or secondary, as appropriate) will record all data related to the resolution check. These data will be recorded as video (and if appropriate, still images). In addition, the analyst (primary or secondary, as appropriate) will fill out the Resolution Check Data Sheet (Attachment B9). The time to perform resolution checks does not count as examination time.
- The examination team will also record data from a resolution standard provided by the test administrator. Data from this auxiliary resolution standard will be used to determine if there is any degradation of image quality in recorded data.
- The test administrator will provide a standard procedure for recording data on the auxiliary resolution standard.
- Supplemental lighting will be permitted as specified in the augmented EP.
- The Analysts will have the freedom to adjust lighting and the angle and magnification of the camera to the extent defined by the augmented EP. If the camera has pan, tilt, or zoom capabilities, these may be used to the extent defined by the augmented EP.

B.5.5 Primary Examination

- All tests will be conducted in a tank, with a blackout tent used to exclude external (or ambient) lighting. The camera will be mounted on a multi-axis scanner controlled by a joystick. The camera will be scanned over each specimen to inspect the designated region of interest.
- The examination team (Primary Analyst and Operator) will be given 20 minutes for the primary examination of each specimen.
- The examination team shall not change cameras during a single test. Separate tests using a different camera may be performed if previously cleared with the test administrator and if additional time is available.

- All flaw calls in the primary examination will be made using only the live feed from the camera. The Primary Analyst will NOT be allowed to make or change calls by reviewing the recorded data (video or still pictures).

B.5.6 Independent Secondary Review and Re-Examination

- Independent secondary review will be performed by the Secondary Analyst using recorded data from the primary examination and the data sheets from the Primary Analyst. During this phase of the secondary review, the Secondary Analyst may NOT consult with the Primary Analyst.
- Re-examinations of a small set of questionable indications or areas will be permitted. At present, only a maximum of 20 of the questionable indications or specific areas containing questionable indications will be permitted to be re-examined.
- All re-examinations will take place at the end of the primary evaluation process (i.e., when all test specimens have been examined by the Primary Analyst).
- Re-examinations may be conducted jointly by the Primary and Secondary Analysts.
- Re-examinations will be conducted using the same camera, equipment, and Examination Procedure used for the primary examination.
- Final dispositioning of indications identified for re-examination will be performed using live data. Reporting and documentation requirements for re-examination will be the same as those for primary examination.
- During the secondary review process, it is recognized that additional reportable indications may be identified by the Secondary Analyst but must be confirmed using live data. These should be reported on the supplied Supplementary Data Sheet (see Section B.5.7). In cases where the additional reportable indications cannot be confirmed using live data because the number of requested re-examinations exceeds the limit for the test, the indications should be reported on the data sheet as Potential Reportable Indications.

B.5.7 Reporting and Documentation

- If a call is made, the Analyst making the call (if possible) will make a still image of the flaw and identify the region containing the flaws using the measuring scale on the specimen.
- All primary evaluation results will be reported using the standard RV-RRT-3 data reporting form (Attachment B4).
- All secondary evaluation results (including re-examination and final dispositioning information) will be reported using the RV-RRT-3 data reporting form (Attachment B4).
- Copies of all data reporting forms will be provided by the test administrator to the examination team prior to beginning each test.
- In addition to the examination results, any other reporting as required by the EP should also be performed, explained, and presented to the test administrators by the examination team after completion of test plate examination.
- Prior to starting an examination (or re-examination) of a test specimen, the Primary (or Secondary) Analyst should take a still picture (if possible) of the specimen label and save this image.

- The examination team will record video documenting the entire primary examination process for each specimen, as well as any re-examination of each selected questionable indication. The recorded video shall also include footage of the specimen label prior to starting an examination or re-examination.
- Audio commentary should also be recorded to help document the detection and identification process during both the primary examination and any re-examination.
- If possible, the Analyst (primary or secondary, as appropriate) shall take a still picture of each reportable indication (primary examination and re-examination) and save this image.

B.5.8 General

- Data shall be stored on agreed-upon storage media in agreed-upon formats. This will include images and video, as well as any other form of data (written documents, etc.).

Attachments to Appendix B

- B1 – Examination Procedure Summary
- B2 – Supplementary Procedural Guidance
- B3 – Agreement to Keep RV-RRT-3 Test Information Confidential
- B4 – RV-RRT-3 Examination Data Sheets
- B5 – Procedure for Alternate Resolution Standard Data Collection
- B6 – RVT-RRT-3 – Alternate Resolution Standard Data Sheet
- B7 – Coordinates
- B8 – Checklist for Test Administrators
- B9 – Resolution Check Data Sheet

Examination Procedure Summary Attachment B1

Team Code:

Examination Procedure Summary:

(This will be developed by the test administrator.

RVT-RRT-3 – Supplementary Procedure Guidance Attachment B2

Objective: The following procedural supplements are provided with the intent to improve crack detection and characterization based on observations made from the Phase II Round Robin results and subsequent testing. These recommendations should be implemented when performing Phase III Round Robin examinations whenever possible.

Detection:

1. The Examiner should perform an overview of the entire test specimen prior to test examinations and define a scan plan for the weld that will assure proper coverage of the inspection area. The scan plan should include:
   a. The field of view (FOV) to be used for scanning
   b. The number of passes for complete coverage
      i. Minimum of one exclusive camera pass per weld depending on weld width
      ii. Minimum of one exclusive camera pass for each heat affected zone
      iii. May require a separate camera pass focused on the weld toe and adjacent HAZ. -Use of increased magnification/decrease FOV on the transition of the weld crown and heat affected zone (HAZ) to assist with the detection of small deviations of the crack path from the weld toe geometry into the HAZ.
   c. Scan type (continuous or stop-and-go)
      i. Scan speed should stop-and-go for areas where surface features are numerous.
2. The Examiner may utilize auxiliary lighting (non-camera mounted) when evaluating possible indications if its use would assist in eliminating shadowing or improving the resolution of small surface features such as texture, weld bead ripple, etc.
   a. Auxiliary lighting is a separate light source not typically attached to the camera.
   b. Auxiliary lighting is typically a diffuse light source providing additional light to the general inspection area.
   c. When used, auxiliary lighting should be positioned as needed to illuminate the surface at an angle not provided by the camera lighting.

Indication Characterization:

1. Following the detection of an indication, the Examiner should maximize the optical resolution taking into consideration; cleanliness of the surface, magnification, angle and lighting and determine indication relevance based on, but not limited to the following considerations:
   - Overall Geometry - IGSCC often exhibits a meandering path and may contain branching. Scratches will typically portray a more linear pattern in comparison. Also, ISSCC may exhibit variable opening widths (e.g. grain fall out areas) while scratches would be expected to have more of a uniform width.
   - End-Point Geometry – IGSCC end points normally taper down, a feature typically not found in scratches.
   - Branching-IGSCC may have branching but scratches typically do not.

- Reflectivity – Crack geometry traps reflective light and will often appear dark. Surface scratches commonly reflect light off the bottom surface when illuminated at the correct intensity and angle.
- Out of-Plane Displacement – With an appropriately placed light source (on-camera or auxiliary lighting), IGSCC may cast a shadow due to a slight out-of-plane displacement where one side of the crack may be slightly higher in elevation than the opposing side.
- Surface Texture – IGSCC typically propagates continuously and will run through grinding marks regardless of depth whereas scratches tend to appear broken (non-continuous).
- Two IGSCC cracks would be less likely to intersect but scratches may.

2. Additional magnification above that used during detection scanning should be used to effectively discriminate between a non-relevant indication and a surface crack.
3. The Examiner should consider deploying different camera angles and lighting options (e.g., auxiliary lighting) in the determination of the relevance of an indication.
4. If the Examiner fails to come to a final conclusion on the relevance of a crack-like indication, the indication shall be evaluated to be a crack.

Agreement to Keep RV-RRT-3 Test Information Confidential

Attachment B3

By my signature below, I confirm that I have read and understand Section B.2.7 (Data Security) of the RV-RRT-3 protocol, and agree to comply with all aspects of the data security requirements.

TEAM SIGN:

Team Member Name (Printed)	Team Member Name (Signature)	Date

ROUND ROBIN DATA SHEET - PHASE 3

Vendor ID: _____

TEST INFORMATION

VT Inspector (Primary): _____

VT Inspector (Secondary): _____

Manipulator Operator: _____

Procedure Number: _____

Date: _____

Notes: _____

SYSTEM INFORMATION (Primary)

Camera ID: _____ ☐ Color ☐ B&W

Zoom Capable: ☐ Yes ☐ No

Camera Resolution: _____

Lighting Type (on camera): ☐ Halogen ☐ LED

Lighting Type (Auxiliary):

Lighting Comments: _____

_____

SRCS Serial Number: _____

SRCS Description:

DATA ACQUISITION SYSTEM INFORMATION (Primary)

Data Acquisition System: _____ _____

Monitor Manufacturer: _____ Model: _____ Resolution: _____ Size: _____

DATA STORAGE:

Recording media used: ☐ Hard drive ☐ DVD ☐ Other

Media Identification Label: _____

COMMENTS:

_____

_____

_____

_____

_____

SIGNATURE: _____ **Date:** _____

ROUND ROBIN DATA SHEET - PHASE 3 (Re-Inspection)

Vendor ID: _____

TEST INFORMATION

VT Inspector (Primary): _____

VT Inspector (Secondary): _____

Manipulator Operator: _____

Procedure Number: _____

Date: _____

Notes: _____

SYSTEM INFORMATION (Re-inspection)

Camera ID: _____ ☐ Color ☐ B&W

Zoom Capable: ☐ Yes ☐ No

Camera Resolution: _____

Lighting Type (on camera): ☐ Halogen ☐ LED

Lighting Type (Auxiliary):

Lighting Comments: _____

SRCS Serial Number: _____

SRCS Description: _____

DATA ACQUISITION SYSTEM INFORMATION (Re-inspection)

Data Acquisition System: _____

Monitor Manufacturer: _____ Model: _____ Resolution: _____ Size: _____

DATA STORAGE:

Recording media used: ☐ Hard drive ☐ DVD ☐ Other

Media Identification Label: _____

COMMENTS:

SIGNATURE: _____ **Date:** _____

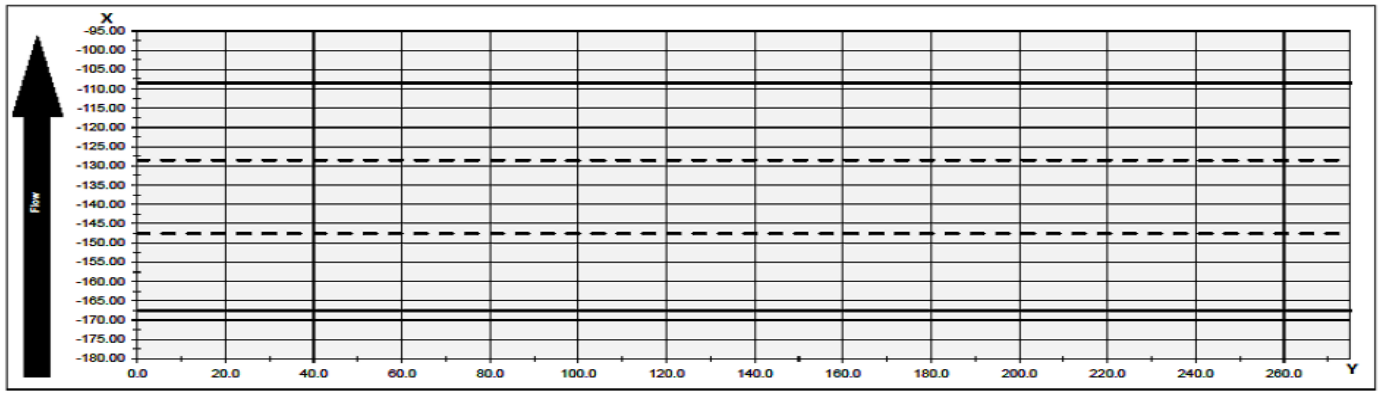
Remote Visual Testing Round Robin Indication Data Sheet

Sample ID: EXAMPLE **Units: mm** **Vendor:** _____ **Date:** _____

Exam Start Time: _____ **Exam End:** _____

Secondary Review Start Time: _____ **Secondary Review End:** _____

Re-inspection Start Time: _____ **Re-inspection End:** _____ **Re-inspection Date:** _____



Candidate Flaw Information

	Analysis Results							Secondary Review	Re-inspection (If conducted)	
Flaw Number	Video Time Stamp	Y start	Y end	X start	X end	Length (mm)	Aux. Lighting?	Decision (Y, N, R)	Final Disposition	Aux. Lighting?

Comments

_____

_____

_____

_____

Analyst Signature: _____

Analyst Signature: _____

Procedure for Alternate Resolution Standard Data Collection — Attachment B5

Objective:

- The objective of this study is to quantify potential degradation in recorded video data (compared to live video), during remote visual examination.
- The objective of this document is to describe the procedure to be used for recording data on alternate resolution standards and for the analysis of these data, in support of the objective of this study.

Equipment: This test will **not** use the standard character standard typically used in RVT. Instead, this test will use an alternate standard (referred to in this document as a resolution standard). While multiple standards may be applicable, one the following is recommended:

- AFRL Resolution Standard
- IEEE Reflection Standard

These standards provide a mechanism for quantifying the minimum discernable line-pair size in both horizontal and vertical directions. The IEEE target also provides for characterizing image degradation using other metrics and/or in other directions (radial).
Important: The target should be physically capable of withstanding submergence without damage.

Procedure Description: This study will be conducted during the initial setup of the systems by RVT-RRT-3 participants.

1. The target will be mounted on/in an appropriate fixture that enables it to be placed underwater, in a plane perpendicular to the camera axis (normal incidence).
2. The RVT camera system will be set up for normal examination. The setup includes the camera and recording equipment, and all necessary parameters such as video encoding parameters, screen resolution, etc.
3. The camera system functionality using the typical RVT examination setup will first be verified (using standard procedures) with the character standard commonly used in RVT examinations. Video data will be recorded from the character standard. If possible, a still picture will also be acquired of the character standard.
4. The camera will placed at a fixed distance (6”) from the resolution standard. The camera zoom function will not be used in this step. The standard will be oriented in such a manner that the horizontal (H) and vertical (V) line pairs align with the H and V axes of the camera (i.e., no relative rotation).
5. Lighting (on-camera only) will be adjusted to maximize the visibility and clarity of the target.
6. A designated person (AKA C. Joffe) will read out and document the finest resolution visible on the screen in both the H and V directions. Depending on the target selected, additional information may be recorded. All information will be documented on the supplied Alternate Resolution Standard data sheet.
7. Video data (and if possible, still pictures) will be recorded at this stand-off distance. The video file with these data will be noted in the data sheet.
8. If the camera system does **not** have zoom capabilities, then go to step 9. If the camera system **has** zoom capabilities, then the camera zoom will be employed to till the edge markings of the resolution target fill in either the horizontal or vertical limits (whichever is reached first) on the monitor. Steps 5-7 will be repeated and the camera zoom reset to a non-zoom setting.
9. Using the scanner, move the camera closer to the target till the edge markings of the resolution target fill in either the horizontal or vertical limits (whichever is reached first) on the monitor (approximately equivalent to the field of view (FOV) with the zoom, in case of cameras with zoom capability). Repeat steps 5-7.
10. If the camera can be rotated about its imaging axis (essentially switching the H and V orientations), then do so, and repeat steps 4-9.

RVT-RRT-3 – Alternate Resolution Standard Data Sheet

Attachment B6

Alternate Resolution Standard Data Sheet

Vendor ID: _____

TEST INFORMATION

Analyst: CHRIS JOFFE

Resolution Standard ID: _____

Orientation: _____

Manipulator Operator: _____

Date: _____

Notes: _____

SYSTEM INFORMATION

Camera ID: _____ ☐ Color ☐ B&W

Zoom Capable: ☐ Yes ☐ No

Camera Resolution: _____

Lighting Type (on camera): ☐ Halogen ☐ LED

Lighting Type (Auxiliary):

Lighting Comments: _____

SRCS Serial Number: _____

SRCS Description: _____

DATA ACQUISITION SYSTEM INFORMATION

Data Acquisition System: _____ _____

Monitor Manufacturer: _____ Model: _____ Resolution: _____ Size: _____

DATA STORAGE:

Recording media used: ☐ Hard drive ☐ DVD ☐ Other

Media Identification Label: _____

Live Analysis:

Minimum Detectable lp: Horizontal: Vertical:

Video File Name: Image File Name:

Recorded Analysis:

Minimum Detectable lp: Horizontal: Vertical:

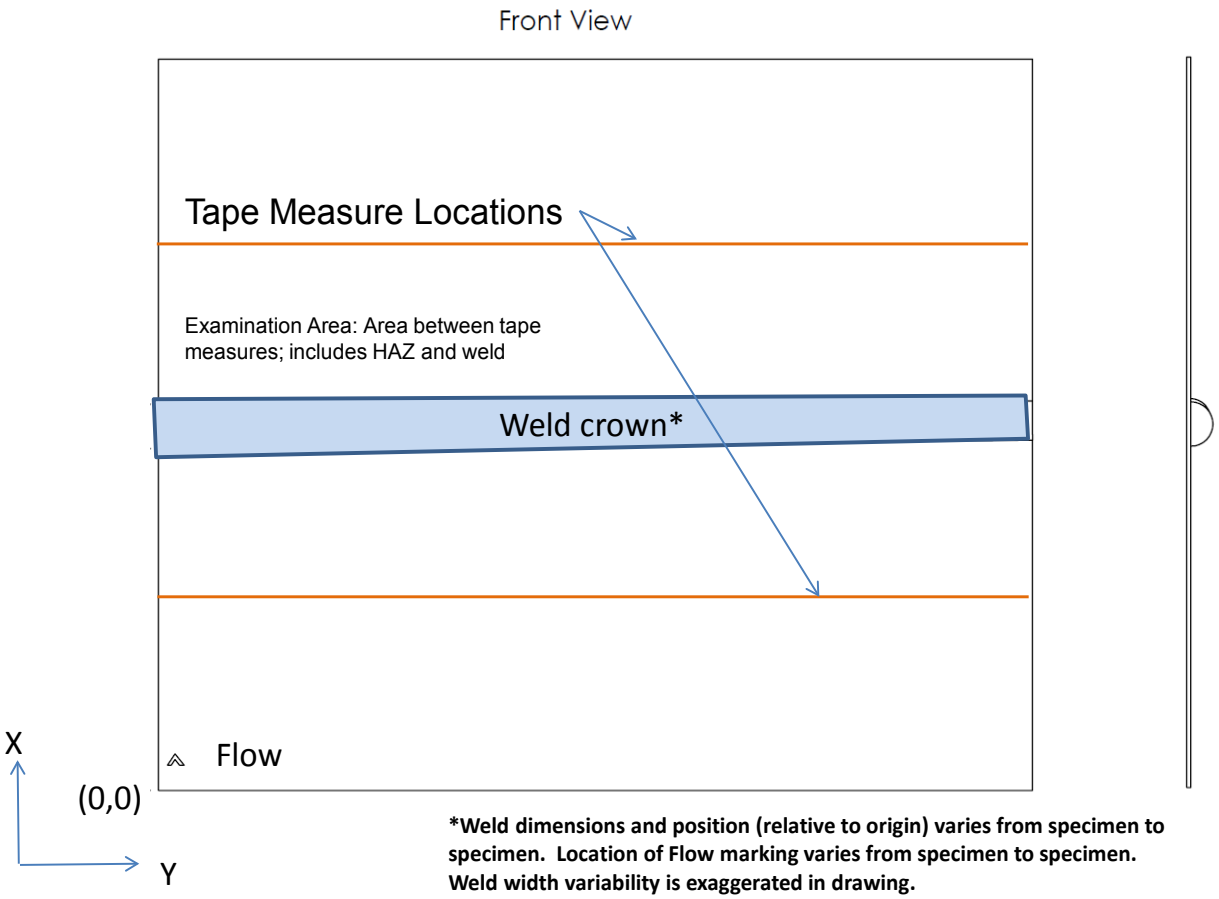
Comments: _____

SIGNATURE: _____ **Date:** _____

Coordinates

Attachment B7

Stainless Steel Specimens



***Weld dimensions and position (relative to origin) varies from specimen to specimen. Location of Flow marking varies from specimen to specimen. Weld width variability is exaggerated in drawing.**

Checklist for Test Administrators Attachment B8

The purpose of this checklist is to facilitate a systematic process for conducting the RV-RRT-3. This checklist presupposes that data acquisition and primary data review are performed at the same time (i.e., using live camera feed), secondary review is performed in parallel with the primary examination, and re-examination is performed after all specimens are inspected once.

Item	Test Administrator Initials	Date	Comments
List of team members			
Test administrator review of test specimens and order of test specimens			
Test administrator observes calibration of equipment and compares to procedure			
Test administrator reviews data acquisition and review, and compares to procedure			
Test administrator reviews inspector fill in of Data Sheets			
Test administrator reviews secondary analysis and fill-in of Data sheets.			
If deviations from procedure are found, make a clear comment and justify deviation, as well as assuring that it is dated and fully documented.			
For re-examinations, Test administrator review of test specimens and order of presentation.			
Test administrator observes calibration of equipment for re-examination and compares to procedure.			
Test administrator reviews data acquisition and review during re-examination, and compares to procedure			
Test administrator reviews inspector fill in of Data Sheets for re-examination.			
If deviations from procedure are found, make a clear comment and justify deviation, as well as assuring that it is dated and fully documented.			

Resolution Check Data Sheet — Attachment B9

SRCS Resolution Data Log				
Vendor:				
	Time	**Date**	**Len-to-Target Distance**	**Comments**
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				
Check - IN				
Check - OUT				

SIGNATURE: _____ **Date:** _____

APPENDIX C CRACK OPENING DISPLACEMENT GROUND TRUTH DETERMINATION

C.1 Overview

This appendix contains a description of the procedure used to locate, measure, and characterize the crack opening displacement (COD) of fatigue cracks used in this visual testing study. This process was done in order to identify true-state of each specimen used, containing both intentional and unintended cracks, to evaluate the visual detection and classification of the cracks from teams using visual inspection techniques. Documented information on each specimen included the weld location, thickness, and orientation with respect to the specimen, as well as location of all detected cracks on the specimen, their length, and the representative-COD value for each crack.

High-resolution images were taken of each flaw to evaluate the length of the crack and COD of the crack along its length. Using MATLAB, measurements along the length in small increments (~2–5 μm, depending on resolution of the image) were made perpendicular to the predominant orientation of the crack. Each set of measured COD values for a crack were then processed to calculate a single COD value that accurately describes the general size of the crack opening as a whole.

Length and COD uncertainty varies from flaw to flaw, largely because of the crack orientation, resolution, and magnification used in the microscopy. Small deviations in crack orientation from the dominant orientation (horizontal or vertical) relative to the microscopy image may result in larger or smaller crack opening measurements at one or more location along the crack. Higher magnifications have a finer pixel-μm resolution. However, any subtle differences in scale across an image set results in a decrease in resolution of each image in that set to match the lowest resolution segment.

C.2 Approach

1. The cracks were imaged using an Olympus BX51 with Stream Motion software. Depending on the length and tortuosity of the crack and the magnification used, several overlapping images may be required to image the crack from end-to-end. Because of limitations on file size (to maintain very high-resolution images), the image sets were often broken up into multiple files, with each file containing several images stitched together by the software. An example of a file with several images already stitched is shown in Figure C-1. Note this is just one of several similar files that constitute the entire crack.

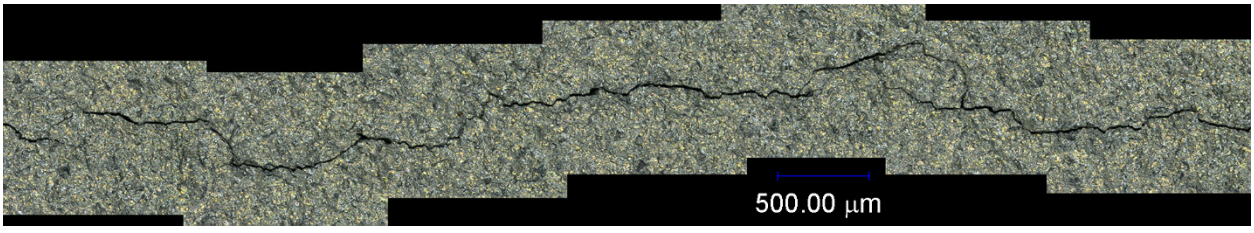


Figure C-1 Crack Segment File with Multiple Photos Stitched Automatically by the Stream Motion Software

2. Highlighting the crack segments in each file and normalizing their size across all the images in the set was done, after collection of the image sets were completed, in Photoshop. While magnification of each crack was generally the same overall, each file had a scale overlaid onto the image (shown in Figure C-1). Because of the high-resolution nature of the image files, the image file resolutions had to be normalized with respect to each other before they could be manually connected. The following steps describe the process used to highlight the crack and normalize the scale across the image set.

   a. Outline crack(s) within each section of the microscopy file set for a crack using the *magic wand* tool in Photoshop (and touching it up manually as necessary). The general wand tolerance used was between 5 and 10 (unit-less). Once satisfied with the crack selection, fill in the region as solid black on a new layer (Figures C-2 and C-3).

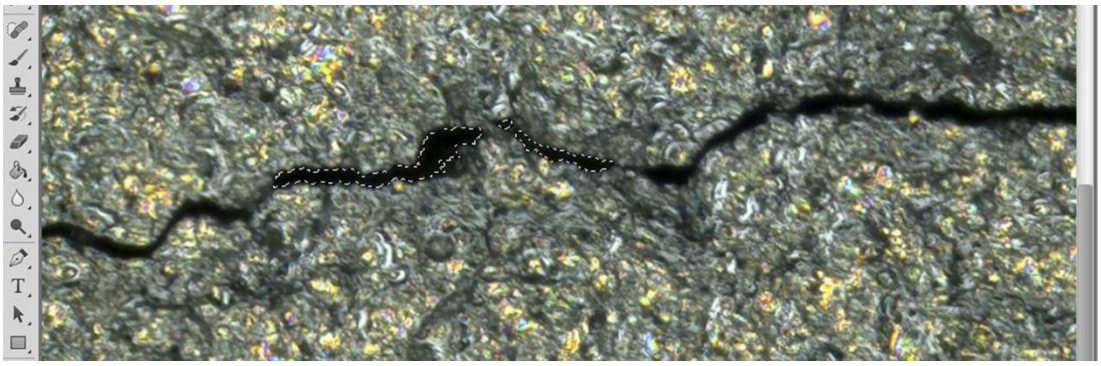


Figure C-2 Selection of the Crack Using the Magic Wand Tool in Photoshop

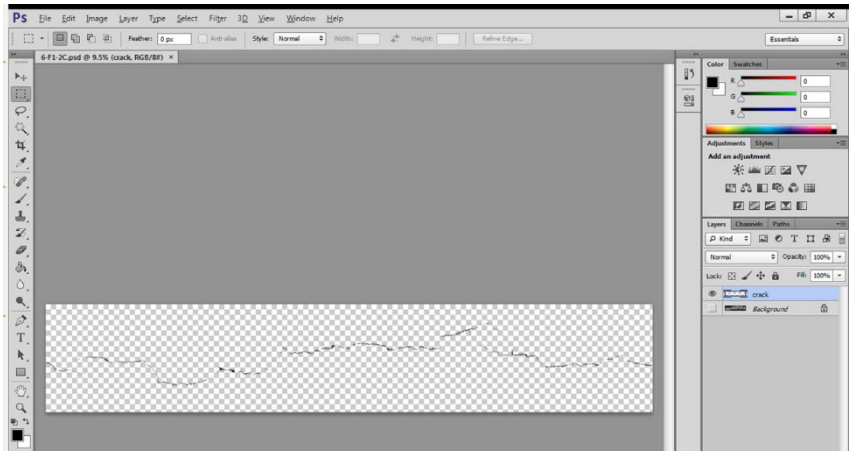


Figure C-3 New Layer Containing a Solid-Black Crack without Anything Else

   b. Normalize the crack image set based on reference length defined from the scales across the images. Each image's scale resolution was calculated for a given set by taking [scale length (px)]/[scale representation (μm)]. The lowest-resolution scale was set as the reference and all of the other images were scaled down (decrease of overall image resolution) to match it. An example of rescaling the images to the lowest resolution can be seen in Table C-1. The lowest-resolution image is segment 2, while the smaller-scaled image (segment 3) is neither the highest nor lowest resolution image in the set. The other two images will be rescaled by the calculated percent ([minimum resolution]/[image resolution]) in both axes to avoid distortion of the image.

Table C-1 Scaling a Set of Crack Images to the Lowest Resolution Across the Set

Crack Segment	Scale, μm*	Scale Size, px	Resolution, px/μm*	% of Original Image Size
1	500	200	0.4	98.0%
2	500	196	0.392	100.0%
3	450	178	0.3956	99.1%

*To convert microns to inches, multiply microns by 0.00004.

3. Manual stitching of the black and white (B&W) crack images was done in Adobe Photoshop and MATLAB. Each crack segment layer created in the previous step was pulled into a new Photoshop file (containing all the other crack segment layers), with each B&W crack segment image contained within its own layer in the new file. The segments were then marked with paired-fiducials so that automatic crack-reconstruction would be simple for MATLAB when loading in the segment images.

   a. Once two crack segments were aligned, the cracks were trimmed to have a flush edge with each other. A red box was drawn with the right side along the seam of those two crack segments. This box was then attached to each crack segment as an identical, fiducial mark for automatic stitching of the later crack-segments (Figure C-4).

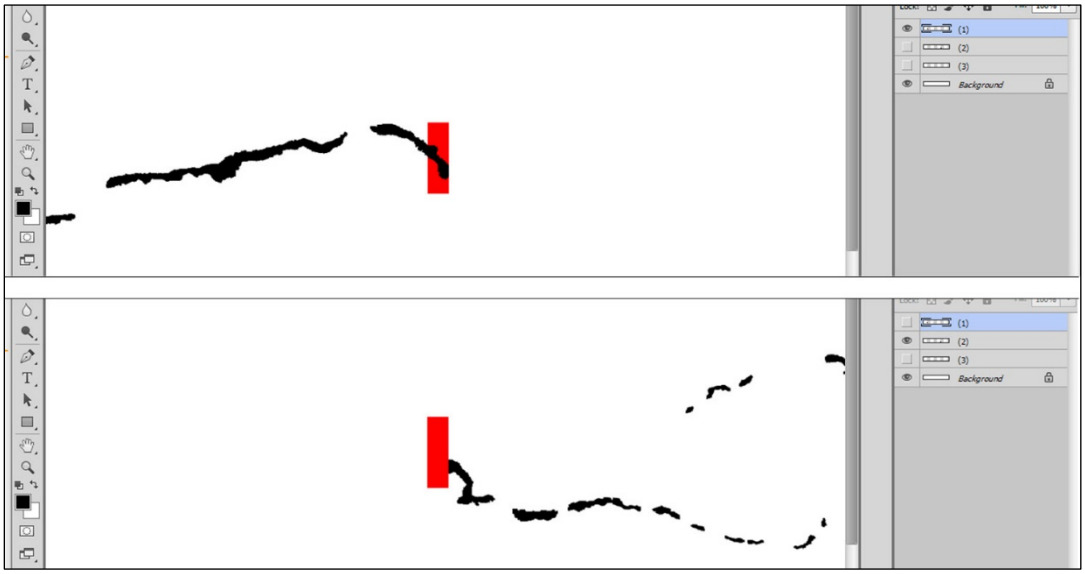


Figure C-4 Example of Two Overlapping Segments of a Crack with the Fiducial Mark (red box) Identified

   b. Center and save each overlapping segment image individually with the attached red box(es) as a jpeg image, numbering the segments from left-to-right (or top-to-bottom).

   c. Using MATLAB, each set of B&W crack segments were incrementally loaded in as color matrices for each crack. By identifying the matching red box pair across each incrementing file, the crack segments were then stitched together as a binary matrix with black representing a “1” and everything else being assigned a “0” (Figure C-5). These crack matrices were then saved as a MATLAB data file for ease in processing.

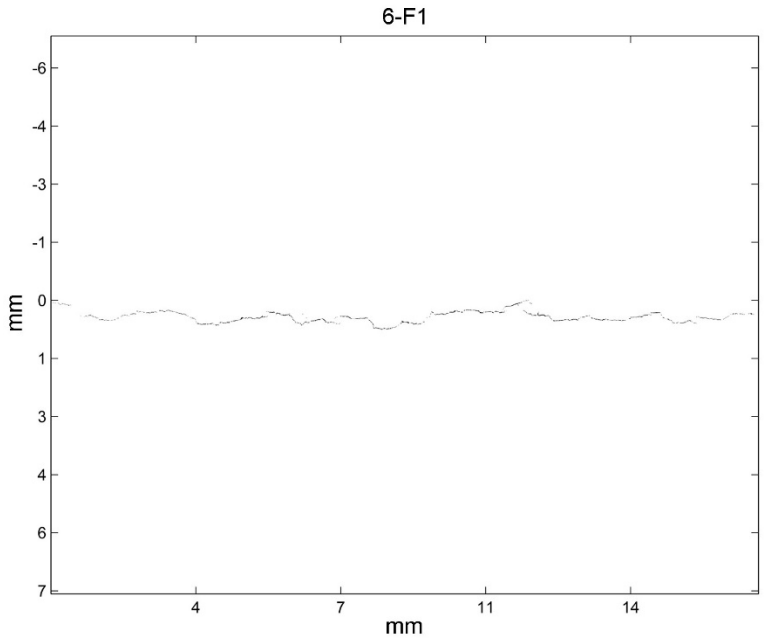


Figure C-5 MATLAB-generated Crack Matrix (from segment images)

4. Further processing of the crack matrices is done to isolate the "dominant" crack segment prior to extracting COD and length information. For a crack with no branching, the dominant segment is the same as the crack. In cases where branching or bifurcation occurs, the ligament that is "closer" to adjacent ligaments is identified as the continuation of that dominant crack. If the separation between ligaments is too large, then they cannot be automatically assumed to be extensions of the same crack and are declared as separate dominant crack segments. MATLAB identifies the dominant crack segments from the ligaments following these rules, and the user then evaluates the segments to define the crack(s) present for length and COD calculations. The steps taken for the final crack definition are summarized below.

   a. Identify each individual segment using MATLAB

      i. Using custom-written scripts, locate and number all of the "holes" in the composite crack image.

      ii. Save all of the segments/branches and their respective COD, gap between adjacent segments (in both horizontal and vertical directions), and mean location along the segment.

   b. Automatically connect segments that are closer than a pre-selected threshold distance (set to 1.5x camera resolution: $1.5\times(125e3/1080) = 174$ μm). This is shown in Figure C-6.

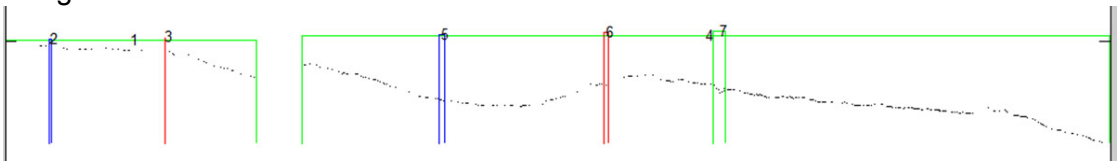


Figure C-6 Automatic Crack Segment Connections

c. Visually verify and identify connected segments. Add/connect other segments manually as needed and approve the connected segments (Figure C-7). The result is an image of the crack from tip to tip.

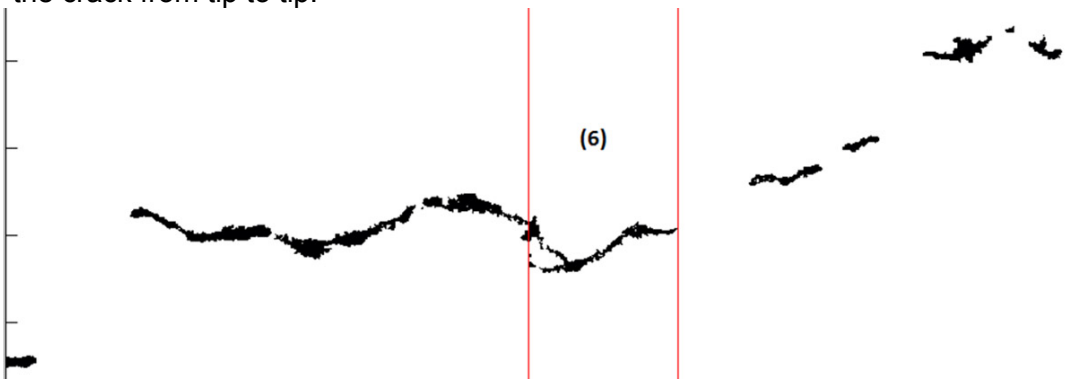


Figure C-7 Zoomed in View on Crack Branch Not Included in Automatic Building

d. Remove all ligaments not declared as being part of the crack from the crack-matrix (Figure C-8). The final crack is then ready to be processed for length and COD true state.

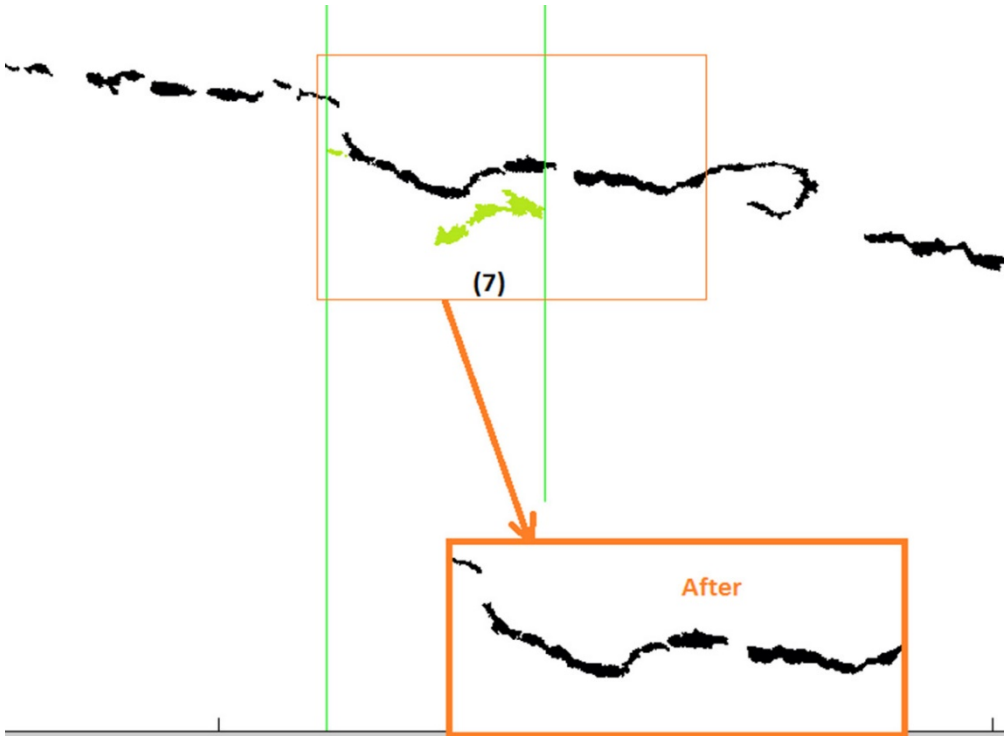


Figure C-8 Example of an Excluded Segment

5. Using the final crack defined by MATLAB, begin automatic measurement of the COD incrementally along the length of the crack. These values are then used to calculate the representative-COD value used to classify the crack along with its length.

   a. From beginning of crack, step through to the end of the crack in 2–5 μm (0.00008–0.00020 in.) increments (based on pixel resolution of the crack image), recording center of crack, COD(s) measured against the predominant crack orientation

(horizontal/vertical), and any gaps that exist if there are multiple CODs. These values are stored in two large matrices (one for the CODs and one for the gaps), and for simplicity of storage any blank location in the matrix is filled with a 0. Table C-2 is an illustration of this.

b. If gaps between parallel segments are smaller than the camera resolution (125e3/1080 = 116 μm), lump those branches together. Define the new COD as the overall distance across the branches (including the respective gaps in between). This is used to avoid small gaps that may only be a few pixels across that arise from measuring the COD across a jagged crack edge, as seen in Figure C-9.

Table C-2 Example of the COD and Gap Matrices

Step Along Crack	COD-1, μm*	COD-2, μm*	COD-3, μm*	Gap-1, μm*	Gap-2, μm*
0	2	0	0	0	0
1	5	0	0	0	0
2	10	0	0	0	0
3	7	3	0	5	0
4	3	2	1	4	2
5	0	0	0	0	0
6	4	0	0	0	0
7	1	0	0	0	0

*To convert microns to inches, multiply microns by 0.00004.

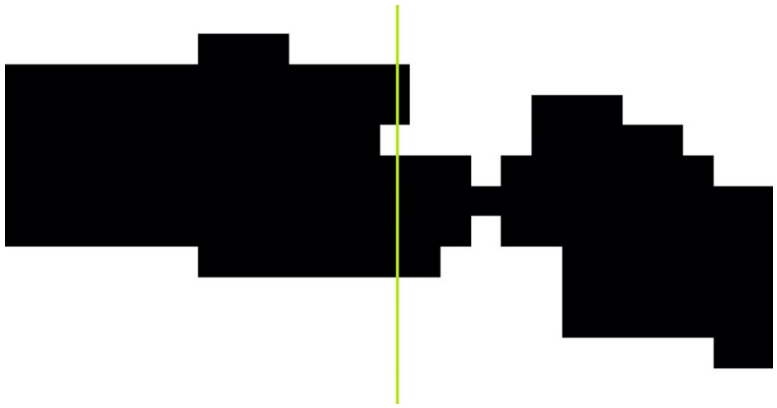


Figure C-9 Small Gap in COD Measurement

6. Calculate statistics on true-state data for the crack, both length and representative-COD

   a. At each position, compute the maximum width across all branches present. If there is only a single branch, the maximum value is simply the width of that branch. From the example in Table C-2, we end up with the following numeric list of the maximum CODs: **2, 5, 10, 7, 3, 0, 4, 1**. Note the zero in the middle of the table (crack is discontinuous or COD is too small to be resolved, etc.).

b. The RMS value of the COD is calculated using the maximum COD values at each location. Using the values in Table C-2 yields:

$$COD_{RMS} = \sqrt{\frac{4 + 25 + 100 + 49 + 9 + 0 + 16 + 1}{8}} = 5.05 \mu m \quad \text{(C.1)}$$

c. Length is defined as the distance between the start and end points of the crack in the predominant orientation.

7. This length value and COD value as determined by COD_{RMS} are then documented as the true-state measurements for the engineering drawings of each specimen.

APPENDIX D POD ANALYSIS PRIMER

NDE inspections can be required to produce two very different types of results. The first type of inspection is called discrete inspection while the second type of inspection is referred to as continuous inspection:

- **Discrete Inspection**: This type of inspection classifies a discrete part into one of two categories: good vs. defective, or flawed vs. unflawed. Examples of such inspections include rivets on an airplane, tube intersections in a steam generator, or castings produced by a manufacturing process. This inspection classifies the entire part into one of the two established categories—good vs. defective or flawed vs. unflawed.
- **Continuous Inspection**: For these inspections, there is a continuous stream of material to be inspected. A prominent example of this type of inspection is the inspection of welds. For such material, flaws occur at a certain rate (rate being defined by length, area, or volume inspected), and the inspection identifies "indications" that hopefully represent actual flaws. A continuous inspection will produce a map of the indications in the inspected material. Some indications will be close to actual flaws and might represent detections; others will be obvious false calls.

From these definitions, we see that it is easy to transform a continuous inspection problem into a discrete one—simply divide the inspected material into units of material of approximately the same size and character. These "grading units" are then treated as discrete parts. Some grading units will contain a flaw and these will be used to determine probability of detection, while others will be blank and be used to determine false call probability.

Because evaluation of discrete inspection is simpler than continuous inspection, the use of grading units is preferred for assessing probabilities of detection and false calls in continuous inspection scenarios. There is an issue as to how to define detection, because a continuous inspection does not explicitly classify a grading unit as flawed/unflawed. The obvious rule to use is that **a grading unit is classified as flawed if the inspection places at least one indication in the grading unit**. From this definition, we see that the grading unit will need to account for a certain amount of sizing error. Also, this definition is used to define a false call—**a false call is a detection in a blank grading unit**.

Note that the definition for detection does not distinguish between an indication that originates from an actual flaw in the grading unit or from other sources in or near the grading unit. As long as one indication, from whatever source, intersects with the grading unit, a detection is considered to have occurred. This perspective is analogous to the "correct answer" grading scheme commonly employed on school tests—a student only has to produce the correct answer.

An alternative definition is obviously that detection occurs only when an inspector produces an indication from **signals originating from the flaw**. If "detection" were defined this way, it would not be possible to have a detection in a blank grading unit. Also, it would be impossible to determine whether or not an inspection produced a detection without evaluating the internal signals produced by the inspection and the decision process employed by the inspector.

This second perspective is analogous to the "show your work" scheme in which the student gets no credit for a correct answer unless s/he shows their work and it produces the correct answer. The "show your work" grading scheme is much more time-consuming to grade, but eliminates the possibility that a correct answer was just a wild guess (equivalent to a false call).

In the study presented in this report, the definition of detection conforms to the "correct answer" scheme; the "show your work" scheme of evaluation for detection is not preferred because 1) it would be very difficult to grade and 2) that is not how detection performance is evaluated for the inspection of a discrete part. Finally, for most inspection procedures, the indication is produced from both signals originating from the flaw and noise in the immediate area of the flaw. Under such circumstances it often becomes impossible to determine whether the indication originated from the flaw or the noise signals.

When transforming continuous inspection to discrete inspection with the use of grading units, it is important to note that the associated detection statistics will not necessarily be independent of each other. If grading units are too close to each other, the detection statistics will be correlated, and POD/FCP estimates will not be described by a binomial distribution.

D.1 Discrete Inspection Performance

For a discrete part, detection performance is typically summarized by a table (Table D-1) that describes the four outcomes possible from inspection.

Table D-1 Possible Outcomes of Discrete Inspection

	Inspection Result	
True State	Unflawed	Flawed
Unflawed	Unflawed call (TN)	False call (FP)
Flawed	Missed flaw (FN)	Detected flaw (TP)

TP = true positive
TN = true negative
FP = false positive
FN = false negative

The results of a round-robin test can be summarized in such a table, with each cell in the table counting the number of grading units that fall into the stated condition. For example, the count in the upper left-hand cell (True Negatives) would represent the number of unflawed grading units that were classified as unflawed by inspection. Direct estimates for FCP and POD are produced from these counts in the obvious way and describe the conditional probabilities of being in each cell (note: PND = probability of no detection) (Table D-2):

Table D-2 Relationship Between Conditional Probabilities from Discrete Inspection

	Inspection Result	
True State	Unflawed	Flawed
Unflawed	1-FCP	FCP
Flawed	1-POD = PND	POD

A more complicated version of the simple categorical table presented above includes flaw size (here described as a percentage but could be in terms of absolute length, depth, COD, or other characteristic of the flaw):

Table D-3 Inspection Result as a Function of Flaw Size

True state Flaw Size (S)	Inspection Result	
	Unflawed	Flawed
$S = 0$	1-POD(0)	POD(0) = FCP
$S = 1\%$	1-POD(1%)	POD(1%)
•		
•		
•		
$S = 100\%$	1-POD(100%)	POD(100%)

When flaw size is included, this becomes a tabular version of a POD curve. If the POD curve is continuous (which it should be), the POD associated with a very small flaw should be equivalent to the FCP. Consequently, the FCP estimate should be used as a data point in any POD fit.

If we use a POD curve to evaluate inspection performance, we see it contains an implicit comparison between FCP and POD. An inspection procedure that is better than guessing must demonstrate that *POD >> FCP*. Typically, an effective procedure must demonstrate that *POD*(S) is large (typically over 80% or 90%, but this is dependent on the specific inspection problem and method being evaluated) for a certain "critical" flaw size.

It should be noted that this flaw detection problem can be formulated as a statistical hypothesis testing problem with the hypotheses defined by:

$\mathbf{H_0}$: No flaw present, $S = 0$

$\mathbf{H_1}$: Flaw of Size $S > 0$ present

In the statistical formulation, the FCP is called the *Type I error*, and POD is equivalent to the probability that H1 is chosen. The POD curve is called the *power curve* and is used to evaluate the test's effectiveness.

One should note than some POD regression models (DoD 2009, see pg. 121 or 129) force POD to be zero when $S = 0$. This is never correct for any real inspection procedure, but might be a reasonable approximation if the FCP is near zero.

D.2 POD Analysis Methodology

Inspection performance is quantified using POD and FCP. Perhaps the best overview of POD/FCP is given by tables that estimate POD/FCP for various conditions. The effect of continuous variables is evaluated using logistic regression models. The most basic model used in these studies is one involving flaw size:

$$\mathrm{POD}(s) = \mathrm{logistic}(\beta_0 + \beta_1 S) \tag{D.1}$$

where S represents flaw size (specifically COD or length in this study) and β_0 and β_1 are the regression model parameters. More complicated models can also be evaluated using logistic

regression. For example, if both COD and length are important, this might lead to the model (with an additional parameter β_2):

$$\mathrm{POD}(\mathrm{COD},\mathrm{Length}) = \mathrm{logistic}(\beta_0 + \beta_1 \mathrm{COD} + \beta_2 \mathrm{Length}) \quad \text{(D.2)}$$

A number of other models are possible and are described in Appendix G.

POD curves can be fit with and without false call data. Most fits in this study used false call data.

The logistic models were fit with the Statistical package "R" using the general linear model (GLM) function, glm() (Hothorn and Everitt 2006; Crawley 2012). The GLM algorithm uses maximum likelihood to determine estimates for unknown model parameters.

To evaluate model goodness-of-fit, the GLM algorithm produces a statistic called dispersion. Dispersion is −2 times the logarithm of the likelihood function divided by the number of model degrees of freedom. This quantity measures dispersion of the binomial data around the fitted curve. If the model fits the data, and only binomial variability is present in the data, the expected value of dispersion is 1. Also when the fit is adequate, the dispersion is distributed as a Chi-squared variable divided by its degrees of freedom.

Frequently, the dispersion statistic will be large, but not due to any systematic misfit of the regression model. A large dispersion statistic will be caused by non-binomial variability in the data. For the inspection data in a typical round robin, such non-binomial variability is caused by flaw-to-flaw (or inspector) characteristics not explained by the regression model. When non-binomial variability exists, a "quasi-binomial" regression model is recommended (McCullagh and Nelder 1983; Agresti 1990). We employ a quasi-binomial model whenever dispersion is greater than 1. It should be noted that the fits of the quasi-binomial and binomial models are the same, but the quasi-binomial model produces different uncertainties and confidence bounds.

D.3 Inspection Grading

To evaluate detection, one must determine when a particular unit of material (blank or flawed) has been called defective. This requires the indications that have been recorded by the inspectors to be associated with the units of material. The units of material being evaluated are called "grading units" and are rectangular units of material, typically containing a flaw in the center.

The inspection team is required to specify the coordinates of an indication that represents a defect as illustrated in Table D-4. Grading associates inspection indications with the grading units we have defined for this analysis.

Table D-4 Example of Indication Data Produced by an Inspection

		Inspection Coordinates			
Indication ID	Inspection ID	Y1	Y2	X1	X2
I3	ALYJ.1	30	50	−153	−153
I4	ALYJ.2	105	125	−125	−125
IT	ALYJ.3	204	206	−125	−116

The following rules are used to make the association and to determine whether or not a unit of material is called defective:

- A unit of material (grading unit) is associated with an inspection indication if both intersect with each other.
- A unit of material is classified as defective when one or more inspection indications are associated (intersect) with the unit.
- POD(S) is the probability that a unit of material, with characteristic S, is classified as defective. The "characteristic S" usually used to distinguish between units is the size of the flaw.

The units of material being evaluated are called grading units, with most units containing a single flaw. However, some grading units are blank and are used to calculate FCP. The result of the grading process is a detection table as illustrated in Table D-5. Each row in the table describes one inspection of one grading unit. If the number of indications that intersect with the grading unit is zero (column 3 in Table D-5), the grading unit was not called defective (i.e., detected), while a value greater than zero in this column signifies a detection. Indications 1 through 3 list up to three intersecting indication IDs.

Table D-5 Example of a Detection Table from Inspection Grading

Inspection ID	Grading Unit ID	Number of Indications Intersecting with Grading unit (N_{det})	Indication 1	Indication 2	Indication 3
1	56	1	2	NA	NA
2	57	1	4	NA	NA
2	58	1	3	NA	NA
12	164	0	NA	NA	NA
6	162	2	16	17	NA

The grading process divides inspection indications into two categories—those that have been associated with a grading unit and those that have not. The properties of the indications that have not been associated with grading units are summarized in a “false call table” as illustrated in Table D-6. Each row in the table describes the false calls for a single inspection. The second column in this table lists the number of unassociated indications in a particular inspection, and the third column represents the length of blank material inspected. These two quantities can be used to calculate an FCR. For example, for inspection 1, we have 1 false call in 256 mm (10 in.) of blank weld inspected, resulting in an FCR of 3.9 per meter (~1 per foot). We can use these statistics to calculate FCRs for various conditions of interest.

Table D-6 Example of a False-Call Table from Inspection Grading

Inspection ID	Number of Un-associated Indications	Inspection Length, mm (in.)
1	1	256.2 (10.09)
2	0	207.6 (8.17)
3	1	239.0 (9.41)
4	1	190.6 (7.51)
5	0	220.7 (8.69)

These FCRs can be related to FCP (for a grading unit of length L) through the Poisson distribution:

$$\mathrm{FCP}\left(\text{GU of length } L\right) = 1 - \exp\left(-\lambda L\right). \tag{D.3}$$

where λ is the FCR for the case of interest and L is the length of the grading unit.

In this study we have utilized three types of grading units: 1) flawed grading units, which contain a single crack; 2) scratched grading units, which contain a scratched area; and clean grading units, which contain no cracks or scratches. Clean grading units are not explicitly identified, because the associated FCP is calculated with the above formula.

Our grading units are rectangles that surround each flaw plus a certain "tolerance." The tolerance is meant to account for an acceptable amount of location error. For the analysis in this report, we have used a tolerance of 10 mm (0.39 in.) in the circumferential direction (parallel to the weld). In other words, we believe that these inspection procedures should be able to locate a flaw to within 10 mm (0.39 in.) in this direction. The tolerance in the direction transverse to the weld varies from 10 mm (0.39 in.) in Phase II to 15 mm (0.6 in.) in Phase III. This was based on an analysis of the impact of the tolerance on the POD, as described below.

Figure D-1 displays the relationship between grading tolerance and POD (using data from Phase II of this study). In ceramic specimens, the POD reaches an asymptote close to 1, while SS seems to be more difficult to inspect, reaching an asymptote of 60%. This analysis indicates that a grading tolerance of 10 mm (0.39 in.) accounts for most of the location errors in the data, and choosing the tolerance to be any higher does not significantly change the POD results.

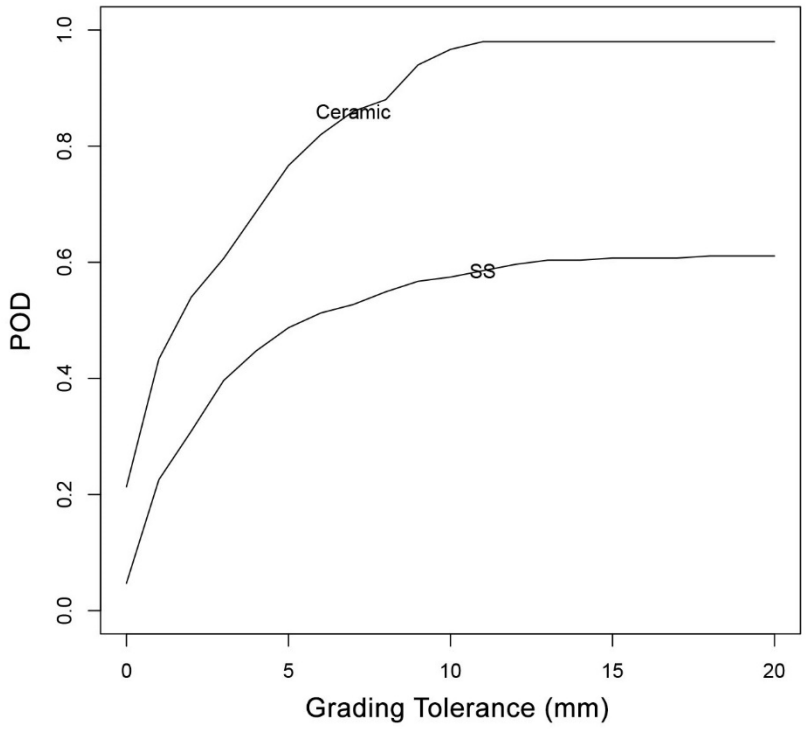


Figure D-1 Plot of POD vs. Grading Tolerance

D.4 False Calls

Two types of false call probabilities are calculated in this analysis. Some portions of the test specimen surface contain surface features—markings on the specimen surface that might be mistaken for a crack (such a scratches, grind marks, etc.). Locations containing surface features (but no flaws) have been identified as surface-feature grading units. These grading units are scored exactly the same as flawed grading units and an associated POD is calculated. In this case, the POD can also be identified as a false call probability.

The remaining material in the specimen (i.e., that material that is not part of a flawed or surface-feature grading unit) is divided into 50 mm (2 in.) long “blank” grading units, and scored for detections. For blank material, an FCR is calculated by counting the number of false calls (i.e., indications not associated with flawed or surface-feature grading units) and dividing by the length of blank material inspected.

APPENDIX E PHASE II RESULTS – DETAILS

This appendix describes the detailed results from Phase II, and supplements the results and descriptions provided in Section 5 of this report. In addition to the detailed analyses presented, some of the tables and figures listed in Section 5 are reproduced here for clarity and additional context.

E.1 Detection Performance Overview

E.1.1 POD Summary Table

The corrected data were scored using a 10 mm (0.39 in.) tolerance, and Table E-1 presents POD estimates on a team-by-specimen–type basis. False call probabilities are also presented for purposes of comparison. One cannot evaluate detection performance using only POD; an effective procedure must exhibit a POD that is significantly larger than FCP. One should note that FCP can be defined in two different ways with this data. First, FCP can be defined as the probability of calling a flaw in a blank grading unit. That is, a grading unit identical to those used for flaws, except it is blank. Second, FCP can be defined as the probability that a surface feature is called a flaw. These two types of FCP are identified as FCP(Blank) and FCP(SF). The table shows that it is much easier to distinguish cracks/blank material as opposed to cracks/ surface features.

Table E-1 provides perhaps the best overview of team performance. First of all, note that the FCR appears to be roughly 1 false call per meter in blank material. This 1 FC/M rate will need to be compared to the rate seen in the field to determine whether or not the inspectors are performing inspections and analyses similar to field inspections. Inspectors have a strong incentive to lower the detection thresholds in a round-robin test so that they can increase their POD scores (but at the risk of increasing FCR). Hence, if observed FCR is much higher than field FCR, one can conclude that this has happened. However, the data for field FCR are difficult to obtain (or indeed quantify) and at present, the information from the RRT should be taken as a baseline, against which subsequent tests/analyses and any information on field performance can be compared to.

Table E-1 POD, FCP, and FCRs by Team and Specimen Type

	Ceramic				Stainless Steel			
	POD	FCP			POD	FCP		
Team	Crack	Surface Feature	Blank	FCR(FC/M)	Crack	Surface Feature	Blank	FCR(FC/M)
ALYJ	1.00	0.40	0.07	1.3E+00	0.78	0.17	0.05	7.9E-01
BMXR	1.00	0.40	0.00	0.0E+00	0.62	0.09	0.02	2.6E-01
CIWN	1.00	0.40	0.00	0.0E+00	0.69	0.05	0.05	7.9E-01
DOYP	1.00	0.40	0.00	0.0E+00	0.64	0.12	0.06	1.0E+00
EQZH	0.83	0.00	0.05	8.3E-01	0.15	0.00	0.02	2.6E-01

The data from Phase II indicate a significant difference in POD and FCR for team EQZH. It is not clear whether this is because team EQZH is using a more stringent detection threshold than the other teams, or for some other reason. The pattern is most noticeable for stainless steel.

One can use the two types of false calls estimated from these data to produce a total FCR (or probability). A formula for a total FCR would have the form:

$$\mathrm{FCR}(\mathrm{Total}) = \mathrm{FCR}(\mathrm{Blank}) + \mathrm{FCP}(\mathrm{SF}) \times \lambda(\mathrm{SF}) \qquad \text{(E.1)}$$

where λ(SF) represents the rate of occurrence of surface features in the material. From the results presented in Table E-1, one can see that total FCR (and therefore total FCP) will be strongly influenced by the rate of occurrence of surface features.

E.1.2 ROC Evaluation of Detection Performance

To allow one to compare team performance visually, we have constructed receiver operating characteristic (ROC) plots of the detection data provided in Table E-1. The two plots shown in Figure E-1 present detection performance in "blank" material and in material with surface features. The individual curves (labelled "C" and "S") show the POD and FCP by individual teams on ceramic ("C") and stainless steel ("S") specimens. An ideal inspection would result in a POD 100% with no false calls. For comparison, the diagonal line on these plots represents the case if one were to simply guess at whether a crack were present or not. As one can see, the presence of surface features makes the detection problem harder. In "blank material" (i.e., material without surface features), detection performance is better in ceramic than SS. However, when discrimination of cracks from surface features is considered, SS and ceramic performance are less distinct.

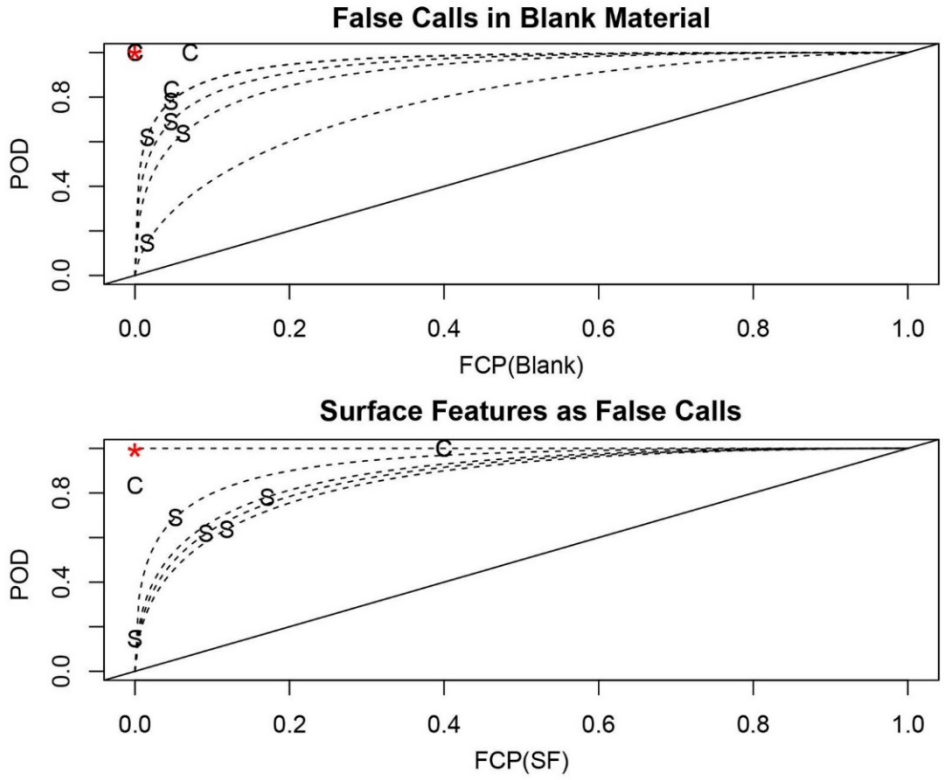


Figure E-1 ROC Plot of (FCP,POD) Estimates in Table E-1. "C": POD and FCP of individual teams on ceramic specimens; "S": POD and FCP of individual teams on stainless steel specimens; *: Ideal inspection system performance.

E.2 POD Analyses

E.2.1 Effect of False Calls on POD Curves

Figure E-2 shows the effect of false calls on the POD regression fit. False call probability is logically equivalent to POD for a crack of size 0 (i.e., POD for a location without any crack), so false call data provide information about the left-hand side of the POD curve. As one can see in Figure E-2, false call data dramatically change the shape of the curve; without the false call data, there is a weak relationship between flaw size and POD.

More specifically, Figure E-2 shows that there is a weak relationship between POD and flaw size in the range of 5 to 80 mm (0.2 to 3.15 in.), but a strong relationship in the 0 to 5 mm (0 to 0.2 in.) range.

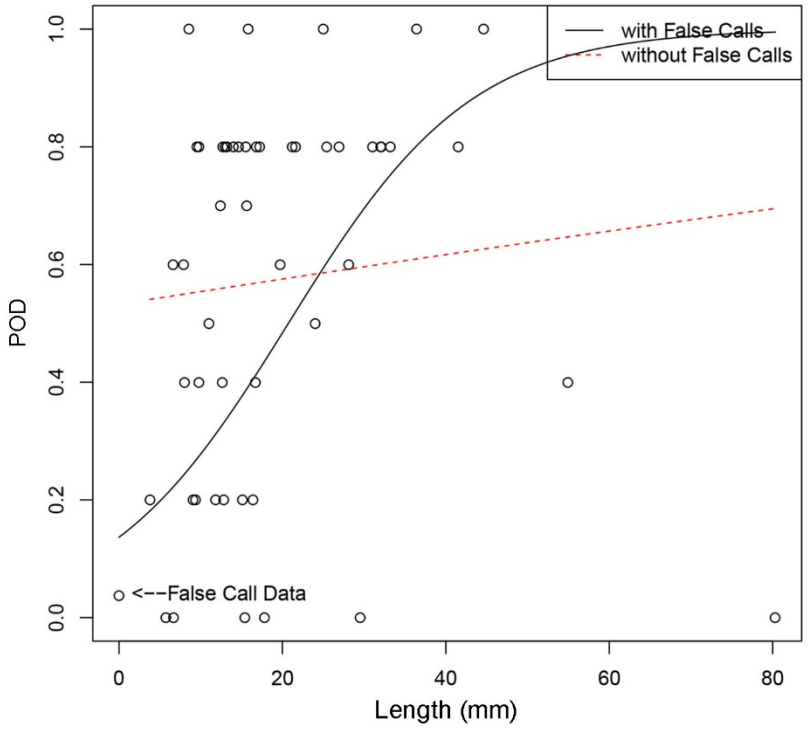


Figure E-2 POD Regression with and without False Calls

For most POD curve fits, we will include false call data and the corresponding curve will be considered to be the most realistic description of in-field POD.

E.2.2 Importance of COD vs. Flaw Length

Which explanatory variable has a stronger effect on POD, COD or flaw length? To answer this question, a logistic model was fit to detection data (without false calls) and with the results plotted in Figure E-3. It appears that only COD affects POD in this data set, and COD is only significant for stainless steel. From the FCPs we have tabulated, we know that both explanatory variables would be important if their ranges were broader.

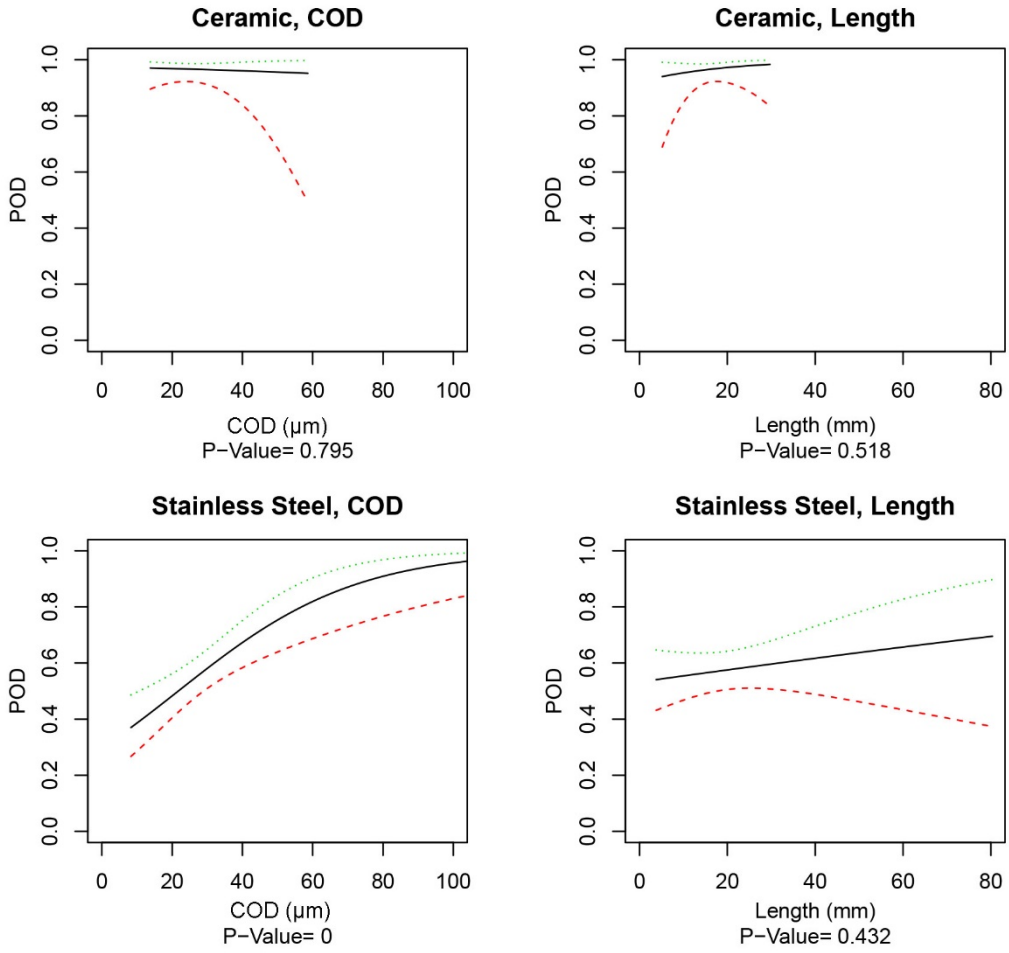


Figure E-3 POD Regression Using the Explanatory Variable COD vs. Flaw Length. The figures show the POD on ceramic specimens as a function of COD (*top left*) and crack length (t*op right*), and in stainless steel specimens as a function of COD (*bottom left*) and crack length (*bottom right*).

Another way to evaluate the importance of POD and flaw length is to include both in a logistic regression model and determine which terms are significant in the regression table. The appropriate regression model is:

$$\mathrm{POD}_i = \mathrm{logistic}\left(\beta_1 + \beta_2 \times \mathrm{COD}_i + \beta_3 \times \mathrm{Length}_i\right) \tag{E.2}$$

and the results of the regression fit for stainless steel are given by Table E-2, which displays the parameters estimates, standard error, and levels of significance. As one can see from the last column in the table, COD is highly significant, while length is not.

Table E-2 Logistic Fit with COD and Length

	Estimate	Std. Error	z Value	Pr(>|z|)
(Intercept)	−0.78	0.31	−2.55	0.01
COD	0.04	0.01	4.31	0.00
Length	−0.01	0.01	−0.59	0.55

The regression fit produces the same conclusions we see in the previous plots.

E.3 POD Curve Using COD and False Calls

This section displays POD curves for COD to be most realistic. Figure E-4 provides an average POD, using the data from all five participating teams. The false call data used in these fits are that produced by "blank" material, not surface features. In the Figure, the top two plots display the estimated curves surrounded by 95% confidence bounds, while the bottom two plots show the fit with data. The bottom two "diagnostic" plots can be used to evaluate how well the regression model fits the data.

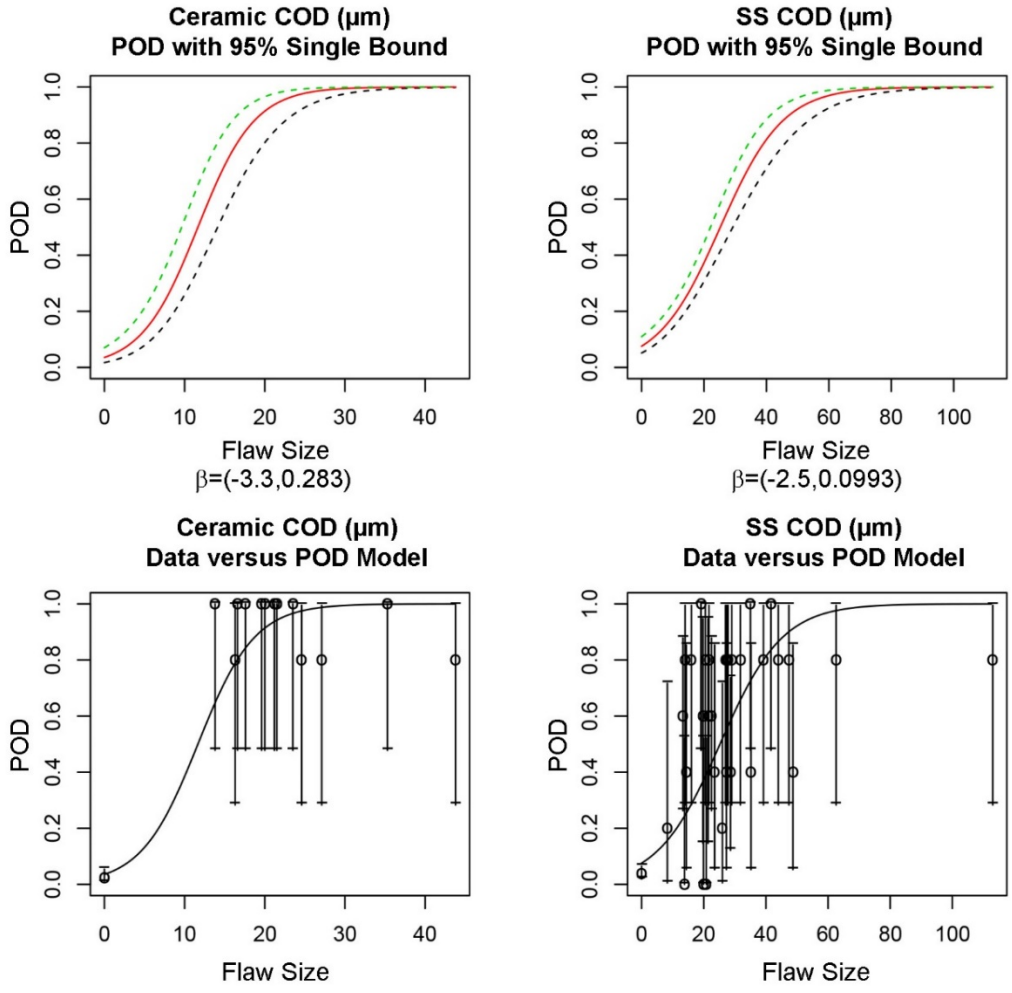


Figure E-4 POD vs. COD

Each data point in these plots represents a single flaw that has been inspected by five teams. The POD point associated with each flaw is surrounded by 95% bounds. If the curve fits the data, it should be within most of these bounds. This is indeed the case.

Figure E-5 presents POD curves for each individual team participating in Phase II. In the ceramic specimens, we can see groups in the POD curves—EQZH forms one group, ALYJ another group, and the other three teams comprise the third group. EQZH exhibits the worst detection performance, while ALYJ the best in the ceramic specimens.

In SS specimens, we see a similar grouping—EQZH forms one group, ALYJ another, and the other three teams form the third. Again, EQZH exhibits the worst detection performance while ALYJ exhibits the best.

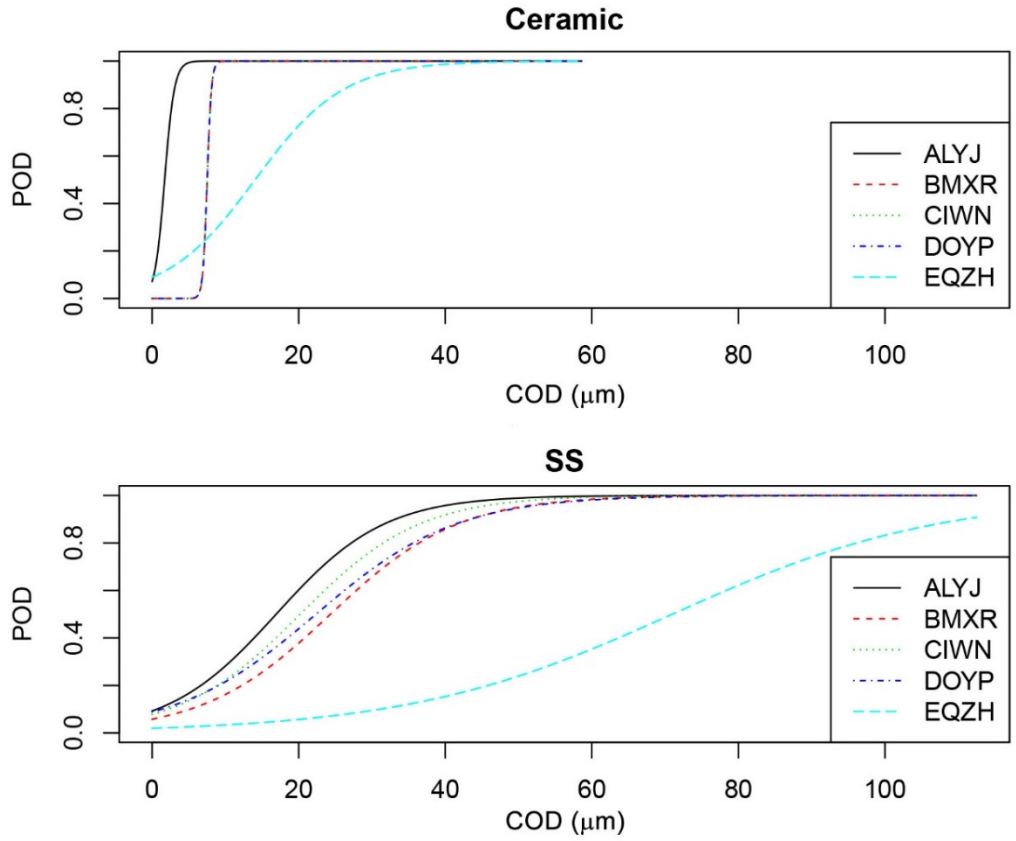


Figure E-5 POD vs. COD for Each Team

E.4 POD Curve Using Length and False Calls

This section produces POD curves using length as the explanatory variable. As noted previously, COD seems to be the better explanatory variable. However a POD vs. length curve is what is most relevant for safety calculations, so these curves are also presented.

The diagnostic plots in Figure E-6 show what is wrong with the POD vs. length regression model; there are two long flaws (length 55 mm [2.165 in.] and 80 mm [3.15 in.]) (not shown on the plot) that cannot be seen. One of these flaws is on the weld edge and very difficult to see. The variable length is not as directly related to visual detectability as flaw width (COD).

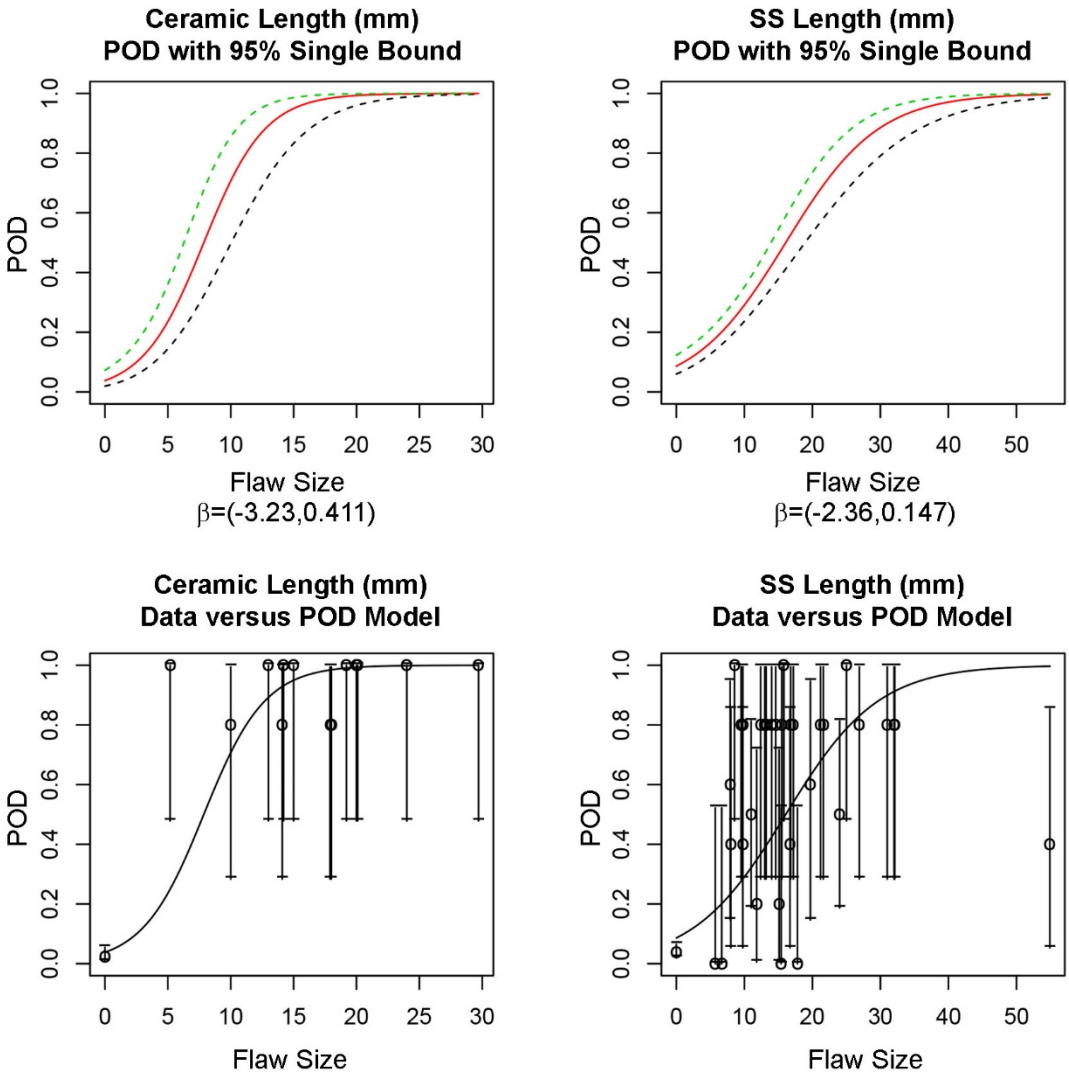


Figure E-6 POD vs. Length Using All Detections and False Calls in Clean Material. Flaw F29.1 excluded.

E.5 Effect of Flaw Explanatory Variables on POD

Tables E-3 through E-5 present the relationship between POD and three explanatory variables—Flaw Orientation, Flaw Location, and Flaw "OnEdge." Because POD is essentially 1 for all flaws in ceramic, we will only examine these explanatory variables in SS. Table E-3 shows a relationship between orientation and POD. However, orientation and flaw location are related to each other. It seems likely that the relationship present in Table E-3 is actually due to flaw location as illustrated in Table E-4; POD in the location HAZ is higher than the other locations.

Table E-3 POD of Axial vs. Circumferential Flaws in SS

Team	Axial	Circumferential
ALYJ	0.69	0.82
BMXR	0.44	0.69
CIWN	0.69	0.69
DOYP	0.62	0.64
EQZH	0.00	0.21
NOBS	16.00	39.00

Table E-4 POD in SS of Flaws in Different Locations: in Ground Weld, in Not Ground Weld, in HAZ, in Surface Feature

Team	Ground Weld	HAZ	Surface Feature	Not Ground Weld
ALYJ	0.67	0.83	0.84	0.57
BMXR	0.33	0.74	0.58	0.57
CIWN	0.67	0.74	0.63	0.71
DOYP	0.50	0.74	0.53	0.71
EQZH	0.00	0.17	0.21	0.00
NOBS	6.00	23.00	19.00	7.00

It was noticed that two very large circumferential flaws placed on the weld edge were not detected at all, and this motivated the construction of Table E-5. From Table E-5, we see that flaws right on the weld edge are indeed difficult to see.

Table E-5 POD in SS of Circumferential Flaws on Weld Edge

Team	Not on Edge	On Edge
ALYJ	0.88	0.57
BMXR	0.75	0.43
CIWN	0.75	0.43
DOYP	0.72	0.29
EQZH	0.25	0.00
NOBS	32.00	7.00

APPENDIX F PHASE III RESULTS – DETAILED ANALYSIS

This appendix describes the detailed results from Phase III, and supplements the results and descriptions provided in Section 6 of this report. In addition to the detailed analyses presented, some of the tables and figures listed in Section 6 are reproduced here for clarity and additional context.

F.1 Detection Performance Overview

F.1.1 Evaluation of Disposition

Because the Phase III inspections contain information concerning the primary and secondary indication dispositions, it is possible to construct tables that describe the primary/secondary decision procedure. Such tables would allow one to determine the contribution of Secondary vs. Primary to detection performance.

Table F-1 has been constructed to allow one to contrast the performance of primary/secondary inspectors. To create this table we tried to associate each indication with one of the three types of material in a specimen—crack material, SF, or blank material. A correct disposition depends on the type of material the indication is in; an effective decision process will have placed all the (GU=Blank or SF) indications in the disp="N," causing false calls to be zero. For the GU=crack category, all the indications should fall into the "Y" category, resulting in the highest possible detection probability.

We can see the disposition procedure is effective. Most of the GU=crack indications are indeed classified as Final.disp=Y (96%), while 50% of the GU=blank indications are classified with Final.disp=Y. For SF material, the final disposition classifies 55% as cracked. So this procedure reduces the FCR/FCP by about 50%, at the expense of reducing POD by 4%. From these results one can see that there may be an opportunity to improve performance by altering the decision procedure so that more indications are identified as false calls.

One can evaluate the secondary's contribution to crack detection by examining Table F-1 line by line. The first line in the table identifies those indications that the primary classified as "NotCrack." To improve the results, the secondary should have overruled the primary for GU=crack indications, but confirmed for both blank and SF. One can see that the secondary did a fairly good job in this role; he made only two "mistakes" (1 crack as N, and 1 SF as crack).

In the second row, which represents those indications that the primary requested a review by the secondary, we find that the secondary again made two mis-classifications (1 blank as crack, and 1 crack as NotCrack). In the final row, the secondary has mis-classified a total of 38 indications. It appears that the secondary may be able to improve his performance most by evaluating the primary's disp=Y calls more critically.

Table F-1 Disposition by Grading Unit Type

Initial Disposition	Final Disposition					
	GU = Blank		GU = Crack		GU = SF	
	N	Y	N	Y	N	Y
N	2	0	1	5	3	1
R	4	1	1	1	1	0
Y	12	17	6	273	9	15

Table F-2 presents recording and detection statistics using all teams. If the disposition procedure employed by the inspection teams was perfect, we would expect to see POR(Crack)=POD(Crack) and POD(SF)=POD(blank)=0. One would classify the disposition procedure as ineffective when

$$\frac{\mathrm{POD(Crack)}}{\mathrm{POR(Crack)}} = \frac{\mathrm{POD(SF)}}{\mathrm{POR(SF)}} = \frac{\mathrm{POD(Blank)}}{\mathrm{POR(Blank)}} \tag{F.1}$$

"Ineffective" in this context means that one could do as well by simply guessing the disposition of each indication. From the perspective of this criterion, the disposition procedure shows some effectiveness. For example, POD(Crack)/POR(Crack) = 97% while POD(SF)/POR(SF) = 77%, demonstrating that POD is reduced only by 3%, while false calls (in SF) are reduced by 23%. In blank material, false calls are reduced by 40%.

Table F-2 Recording and Detection Statistics for Types of GUs

	NOBS	Nrec	POR	Ndetp	PODP	NdetF	PODF
Bonus	40	21	0.53	18	0.45	19	0.47
Crack	375	298	0.79	290	0.77	290	0.77
SF	490	65	0.13	59	0.12	49	0.10
Blank	635	34	0.05	27	0.04	16	0.03
NOBS = number of observations							

Table F-2 presents recording probability (identified as POR), the primary probability of detection (identified as PODP), and the final probability of detection (identified simply as PODF). The statistics in the "SF" and "Blank" rows actually represent false call probabilities. From this table, one can see that the recording step is fairly effective in detection (POR(crack)=73% and POR(blank)=FRP=8%). The main effect of the disposition procedure is to reduce FCP by 25%.

We can see that the secondary's principal contribution to detection performance lies in the reduction of false calls. Of course, one must observe that the primary and secondary inspectors are not making decisions entirely independently of each other. If the primary knew his decision was to be the final decision, he might have classified more indications as "not crack," and thus reduced his FCP more dramatically.

F.1.2 POD/FC Summary Tables

The data were scored using a (10 mm, 15 mm; 0.39 in., 0.6 in.) tolerance, with Table F-3 presenting POD results by team. False call probabilities are also presented for purposes of

comparison. It should be emphasized that POD in this section refers to the "final" POD (i.e., PODF as determined by the secondary inspector's disposition).

Table F-3 POD, FCP and FCRs by Team, FCR=(FC/M)

	POD.Crack	FCP.SF	FCP.Blank	FCR
ARLW	0.72	0.11	0.01	1.59E-01
DCSI	0.72	0.06	0.06	1.12E+00
NBIE	0.85	0.13	0.02	3.19E-01
TUQZ	0.80	0.10	0.03	6.38E-01
YPJH	0.77	0.09	0.02	3.19E-01
All Teams	0.77	0.10	0.03	5.10E-01
NOBS	75.00	98.00	127.00	

The three columns most relevant for the evaluation of detection performance are the first three; these are used in the next section to produce ROC curves for each team. The last column, FCR, presents the FCR in units of false calls/meter. FCP.Blank represents a false call probability for a unit of material of length 50 mm (2 in.).

From this table, it seems teams NBIE and TUQZ are the best, achieving the highest POD, while also achieving relatively low FCP. The "All Teams" row presents average performance over teams. On average, POD is 77% with a FCR of 0.51 indications/meter.

F.1.3 ROC Plot for Detection Performance

To allow one to compare team performance visually, we have constructed ROC plots of the detection data provided in Table F-3. The circles show the POD and FCP by individual teams on the stainless steel specimens. An ideal inspection would result in a POD 100% with no false calls. For comparison, the diagonal line on these plots represents the case if one were to simply guess at whether a crack were present or not. The two plots shown in Figure F-1 present detection vs. false call performance in "blank" material and in material with surface features. As one can see, the presence of surface features makes the detection problem harder.

From these plots, we can see that all five teams have approximately the same detection performance when surface features are used to represent false calls. If welds have many "surface features" and it is necessary for the inspections to distinguish between surface features and flaws, then the second ROC plot in Figure F-1 is the relevant ROC plot to use. In "clean" welds, there is a difference in performance, with NBIE and TUQZ showing the best detection performance, and DCSI markedly worse performance than the other teams.

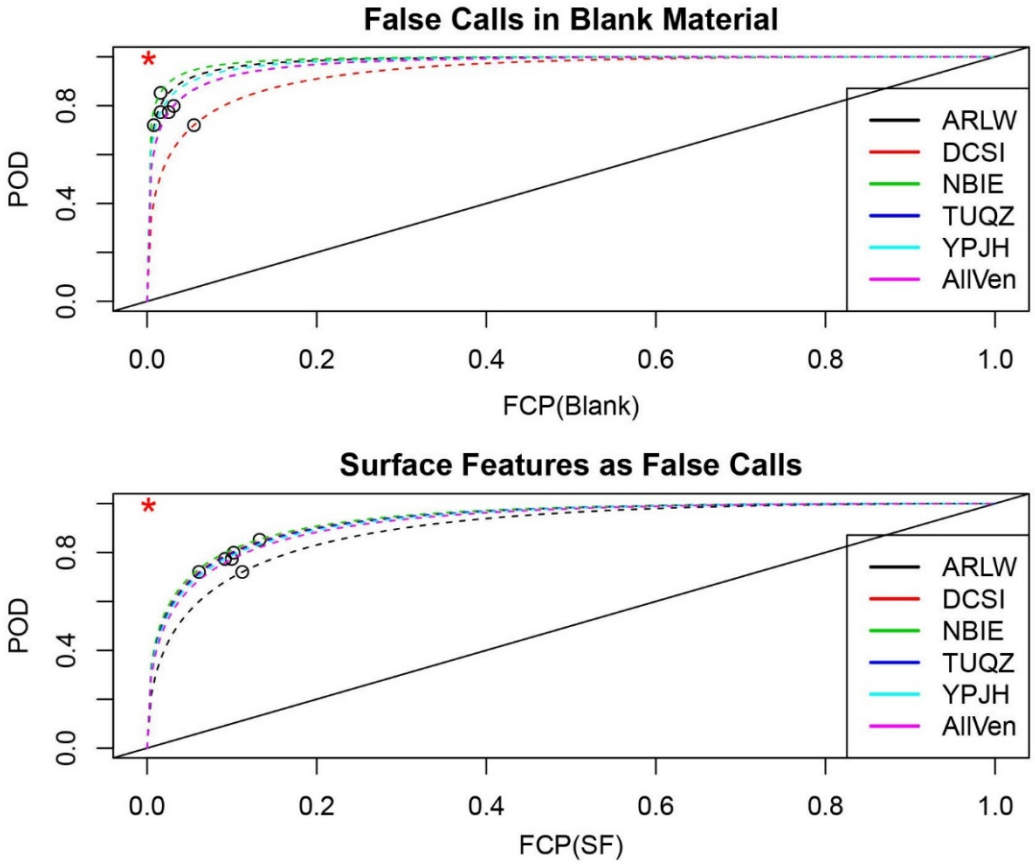


Figure F-1 ROC Plot of (FCP,POD) Estimates in Table F-3. "o": POD and FCP of individual teams on stainless steel specimens; *: Ideal inspection system performance.

F.2 Detection Performance

F.2.1 Evaluation of POD Curve Models

F.2.1.1 Effect of False Calls on POD Models

Figure F-2 shows the effect of false calls on the POD regression curve. False call probability is logically equivalent to POD for a crack of size 0 (i.e., POD for a location without any crack), so false call data provide information about the left-hand side of the POD curve. As one can see in Figure F-2, false call data dramatically change the shape of the curve; without the false call data, there is a weak relationship between flaw size and POD.

More specifically, Figure F-2 shows that there is a weak relationship between POD and flaw size in the range of 5 mm to 60 mm (0.2 in. to 2.36 in.), but a strong relationship in the 0 mm to 5 mm (0 in. to 0.2 in.) range. The most dramatic change in POD seems to be occurring in the 0–5 mm (0–0.2 in.) interval—an interval we would have no data about if the false call data were eliminated. This, incidentally, is the same relationship we saw in the Phase II study.

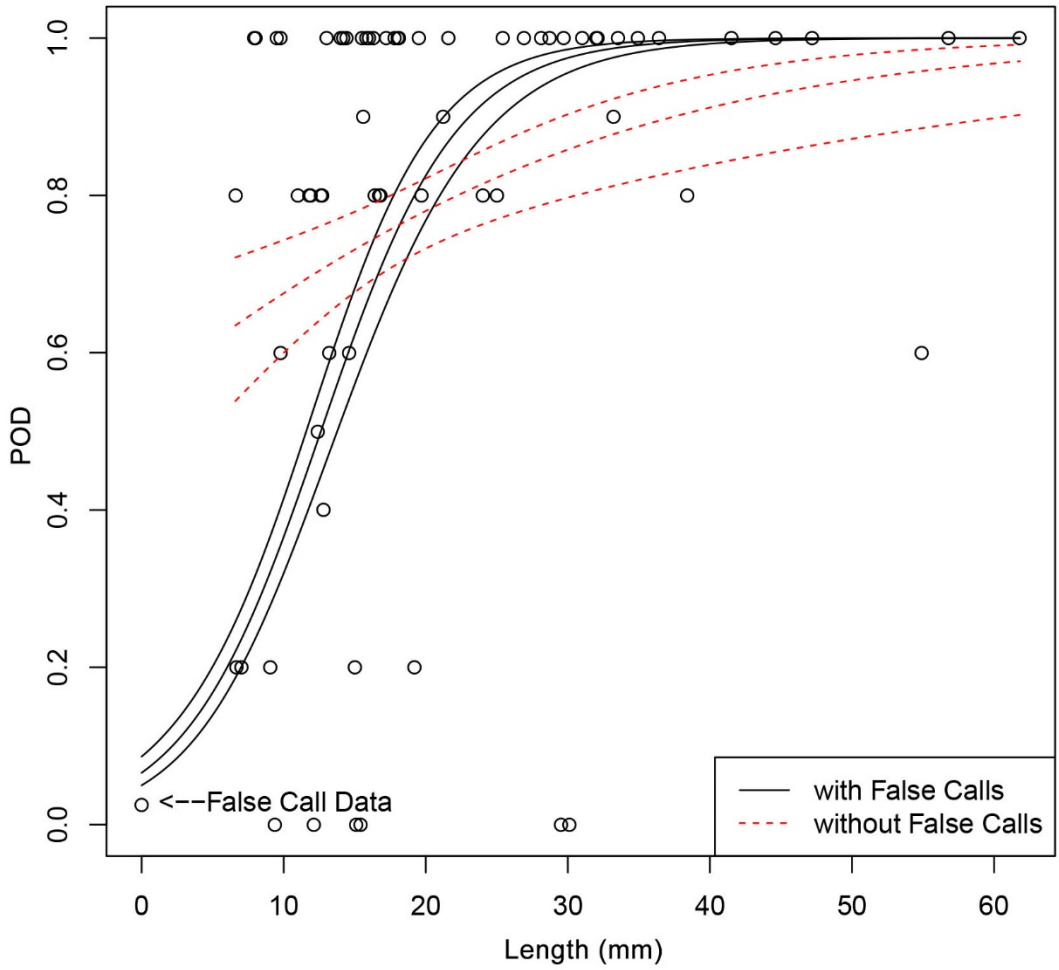


Figure F-2 POD Regression with and without False Calls

The most complete description of POD vs. flaw size is produced when false calls are included, so almost all POD model fits in this report will include false calls. The false call probability included in the fits is that associated with blank material.

Inclusion of false call data requires that the POD model cannot be constrained to be 0 at POD(0). Thus a POD model such as

$$\mathrm{POD}(s) = \mathrm{logistic}\left(\beta_0 + \beta_1 \log(s)\right), \tag{F.2}$$

where s is the independent parameter (length or COD), and β_0 and β_1 are the unknown model parameters, cannot be fit with the false call data included.

F.2.1.2 Evaluation of Possible POD Models

Which explanatory variable has a stronger effect on POD, COD or flaw length? And which POD model involving these variables fit the data the best? To evaluate the model fit, Table F-4 provides goodness-of-fit (GOF) statistics for seven plausible models. In order to calculate GOF for the models, the team inspections were treated as replicates, an assumption that is reasonable for

these data. The GOF statistic presented in the table is “dispersion,” and as the name implies, quantifies the scatter of the data around the POD curve. A “large” value for dispersion indicates model mis-fit and a value near one would indicate a “good” fit to the data.

Table F-4 Summary of Various POD Model Fits to Data

	Model	DOF	Goodness of Fit (dispersion)
1	POD = logistic ($\beta_1 + \beta_2$ COD)	116	1.97
2	POD = logistic ($\beta_1 + \beta_2$ Length)	116	3.23
3	POD = logistic ($\beta_1 + \beta_2$ Area)	116	2.49
4	POD = logistic ($\beta_1 + \beta_2 \sqrt{}$Area)	116	1.91
5	POD = logistic ($\beta_1 + \beta_2$ COD + β_3 Length)	115	1.81
6	POD = logistic ($\beta_1 + \beta_2$ log10 (COD))	73	2.41
7	POD = logistic ($\beta_1 + \beta_2$ log10 (Length))	73	3.11

Note: Models 6 and 7 do not use FC data.
DOF = degrees of freedom

In Table F-4, the model that produced the best fit to the data was Model 5, the model that included both length and COD as linear factors, while the worst was Model 2, the model that utilized length alone. Even though Model 5 exhibits the best GOF, it does not make physical sense when Length=0 or COD=0. A more reasonable model is one that involves crack area, or some function of area, as a measure of size. Model 4 provides the GOF for a POD model involving area; as one can see, the GOF is relatively high. An examination of diagnostic plots showed that the POD was too large for “large” flaws. To force the POD-area model to behave more like the POD COD model, Model 4 was considered, which used the square root of area. This produced a fit almost as good as Model 5, but one that behaves reasonably when Length=0 or COD=0. That is, the FCP is produced. From these results, it appears that Model 4 or Model 1 provide the best description of detection performance.

Even though Model 2 fit the data poorly, it is important, because for safety evaluations, one really requires POD as a function of crack length. A few important points about this fit are in order. First of all, the large GOF statistic is not due to deficiencies in the shape of the POD curve, but rather due to individual flaws with large length but small COD. See the diagnostic plot in Figure F-3 and note the two flaws at 30 mm (1.2 in.) that were missed by all five teams, and also the flaw at 55 mm (2.165 in.) that was missed by two-fifths of the teams. Hence, the difference in the GOF between the models involving COD and length is because detectability is more closely related to COD than length.

Models 6 and 7 in the table are included for reference. These models force POD to be zero at COD=0, so no false call data can be included in the fit. However, the GOF of these models is about the same as that for Models 3 and 2, respectively. Thus, Model 6 seems to fit the data about as well as Model 3, while length (Models 2 and 7) appears to correlate poorly with POD for RVT.

F.2.1.3 Comparison of POD Curves Using COD and Length

This section presents the POD models involving COD and length in more detail. POD curves with 95% confidence bounds and diagnostic plots are presented and may be compared to Phase II results.

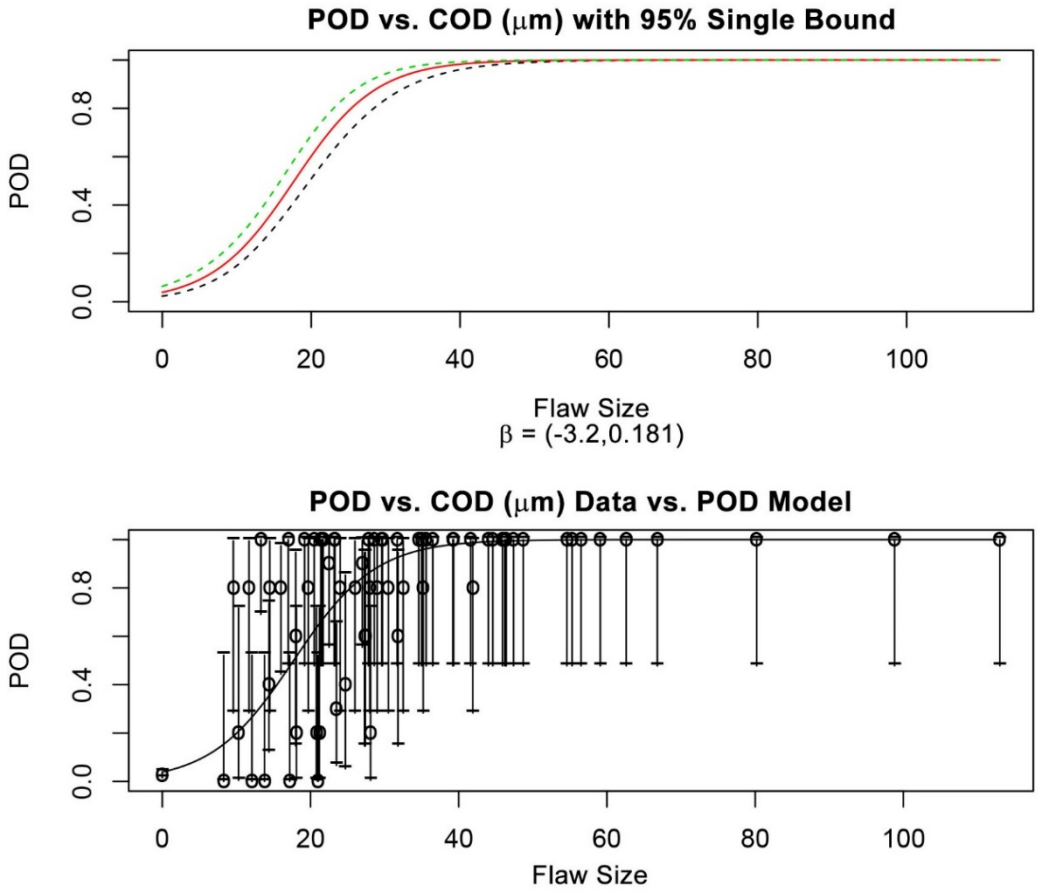


Figure F-3 POD vs. COD

The diagnostic plots presented in Figures F-3 and F-5 show the data vs. the curve fit. Each “data point” represents a flaw, with detection achieved by the five teams. Each data point is surrounded by 95% confidence bounds and a point that does not fit would be indicated by a point whose bounds do not intersect the fitted curve. There are no such points (flaws) in the COD curve fit, but there are such flaws in the length fit. For the length fit, note the flaws at 8, 19, 30, and 55 mm (0.315, 0.75, 1.2, and 2.165 in.). It is these flaws that make the GOF poor for the POD vs. length model.

It is important to note that the “poor” fit of the length model is not due to any deficiency in the regression model form, but due to flaw-to-flaw variations in detectability that cannot be accounted for by the variable flaw length. Therefore, this POD model provides the best description of POD vs. length **for this set of cracks**. If one were interested in calculating a POD vs. length curve for another population of cracks, it would be best to use the curve using the POD vs. COD model (which fits best) and the joint distribution of (COD,Length) assumed in the population.

For example, if we were interested in determining the POD for "field inspections" and knew the conditional distribution for in-field flaws was *f*(COD|Len), the POD(Len) could be calculated from POD(COD) using:

$$\mathrm{POD}(\mathrm{Len}) = \int \mathrm{POD}(\mathrm{COD}) f(\mathrm{COD} \mid \mathrm{Len})\, d\mathrm{COD}. \qquad \text{(F.3)}$$

Individual curves are presented for each team in Figures F-4 and F-6. It appears that most teams produced very similar POD curve performance. The outlier is team DCSI, with a higher FCP and lower POD for large flaws. Because the five teams are so similar, we can conclude the VT inspection protocol and training is achieving consistent performance.

Finally, Tables F-5 through F-8 present the crack size associated with a POD of 80% and 90%. These values are generally accepted target values for acceptable performance. The tables also contain upper and lower bounds for the flaw size estimates. From these tables, we see that 80% POD is reached at COD = 25 microns (0.001 in.) and length = 19 mm (0.75 in.), while the 90% POD is reached at a COD of about 28 microns (0.0011 in.) and length of approximately 23 mm (0.91 in.). The confidence bounds presented here reflect crack sizes for a POD of 80% or 90%, at confidence levels of 2.5% and 97.5%. Similar bounds may be computed to reflect crack sizes at confidence levels of 5% and 95%. From the data presented here, the crack sizes for a POD of 90% at a confidence level of 97.5% (referred to as the 90/97.5 crack size or $a_{90/97.5}$) are seen to be a COD of approximately 33.61 microns (0.00134 in.) and a length of approximately 27.5 mm (1.08 in.).

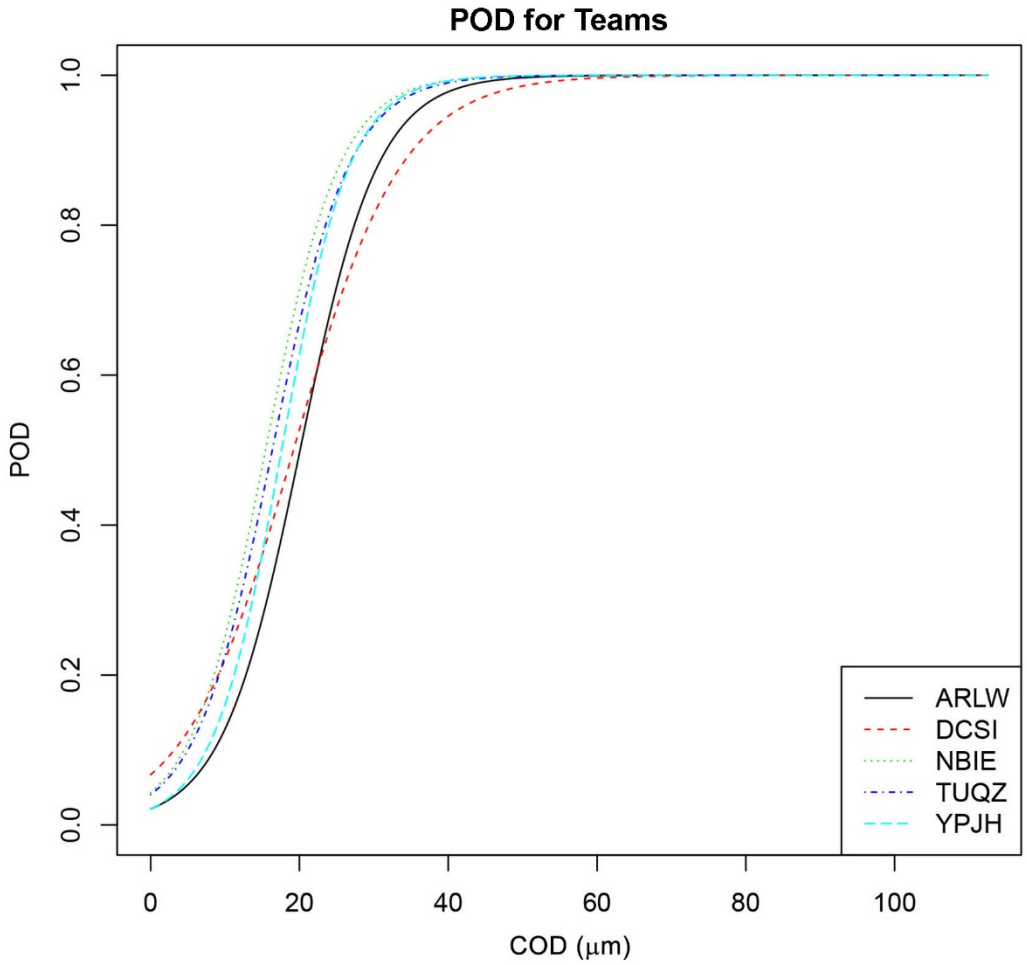


Figure F-4 POD vs. COD for Each Team

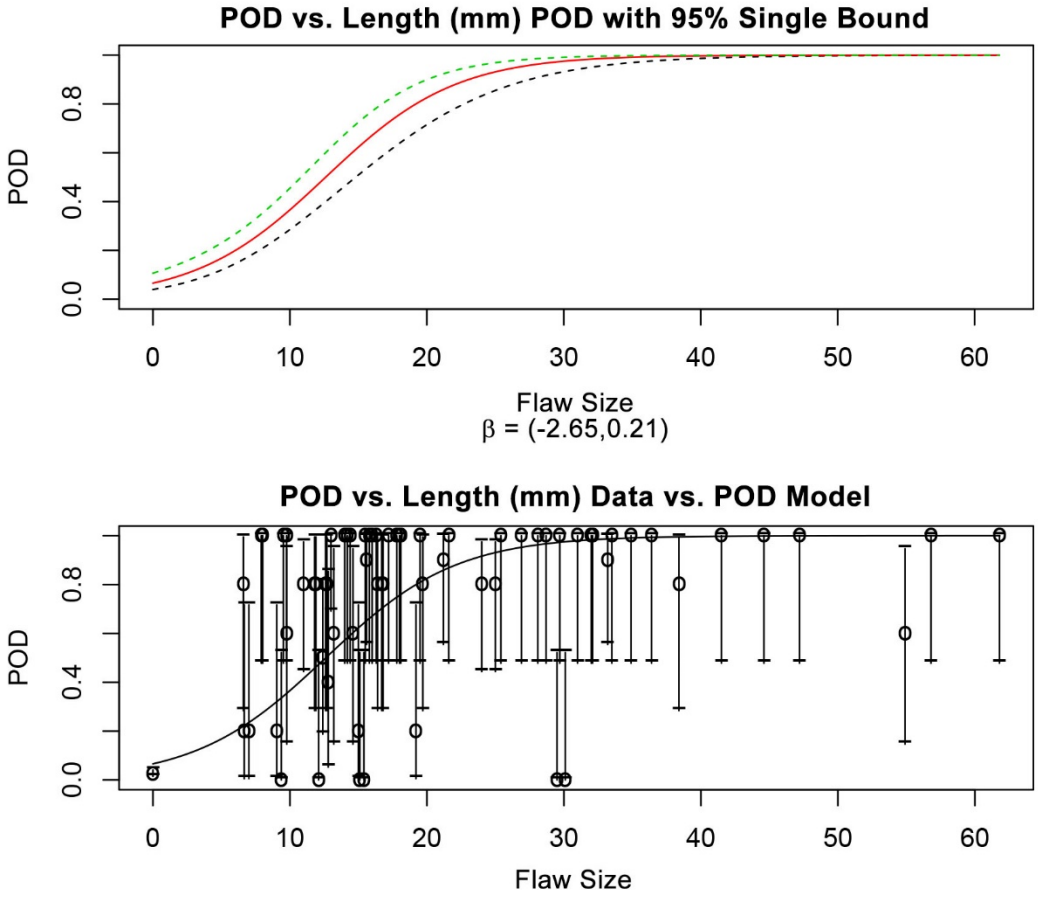


Figure F-5 POD vs. Flaw Length

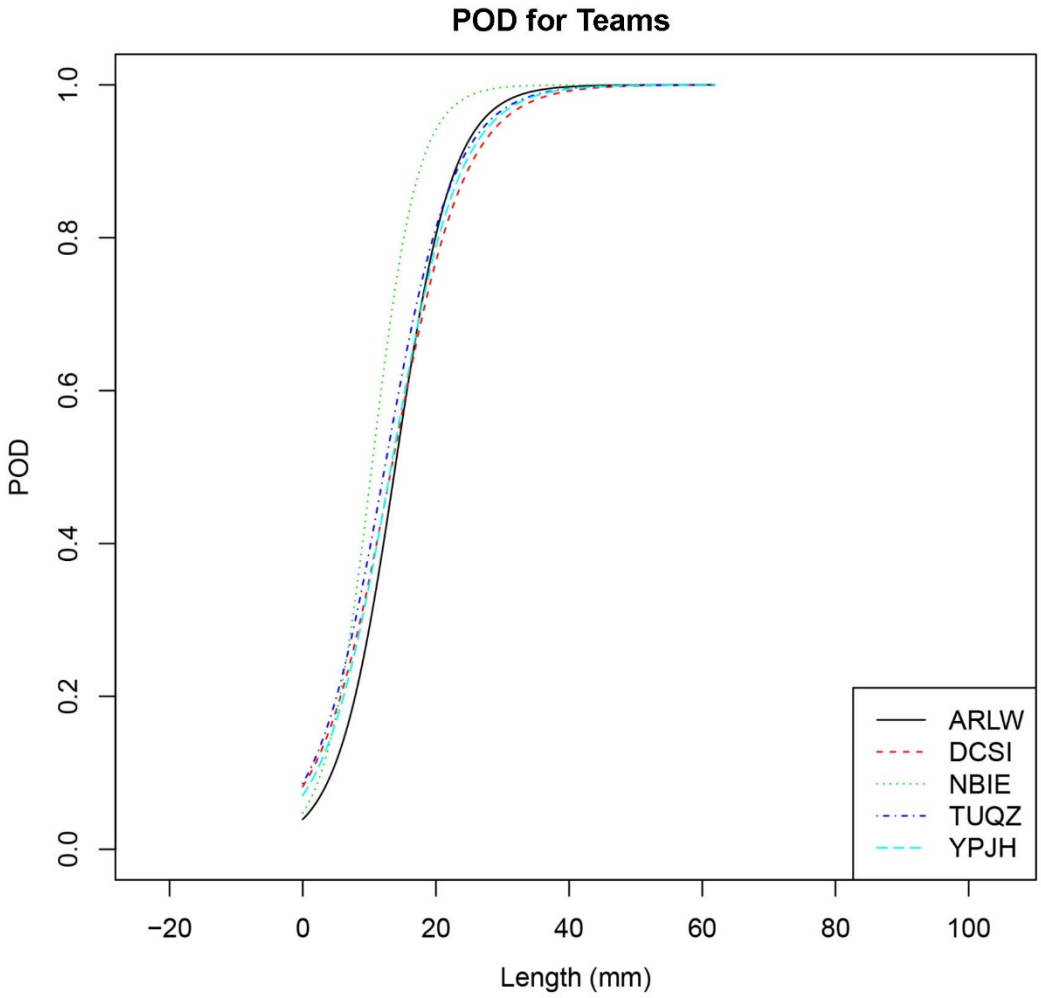


Figure F-6 POD vs. Length for Each Team

Table F-5 Estimate of Crack Size (COD) Associated with 80% POD for Each Team. Bounds are 95%.

Case	Lower Bound, μm*	Flaw Size, μm*	Upper Bound ($a_{90/97.5}$), μm*
ARLW	23.9	27.4	32.6
DCSI	24.8	29.3	36.3
NBIE	19.0	22.4	27.2
TUQZ	20.1	23.5	28.5
YPJH	20.8	24.0	28.6
All	23.0	25.4	28.4

*To convert microns to inches, multiply microns by 0.00004.

Table F-6 Estimate of Crack Size (COD) Associated with 90% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, μm*	Flaw Size, μm*	Upper Bound ($a_{90/97.5}$), μm*
ARLW	27.5	31.6	38.2
DCSI	29.7	35.2	44.0
NBIE	22.8	26.4	32.5
TUQZ	23.7	27.7	33.9
YPJH	24.0	27.7	33.4
All	27.0	29.9	33.6

*To convert microns to inches, multiply microns by 0.00004.

Table F-7 Estimate of Crack Size (Length) Associated with 80% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, mm (in.)	Flaw Size, mm (in.)	Upper Bound ($a_{90/97.5}$), mm (in.)
ARLW	17.1 (0.67)	19.9 (0.78)	24.3 (0.96)
DCSI	17.4 (0.69)	21.0 (0.83)	27.0 (1.06)
NBIE	12.9 (0.51)	15.2 (0.60)	18.5 (0.73)
TUQZ	16.3 (0.64)	19.6 (0.77)	24.9 (0.98)
YPJH	17.1 (0.67)	20.4 (0.80)	25.7 (1.01)
All	16.7 (0.66)	19.2 (0.76)	22.7 (0.89)

Table F-8 Estimate of Crack Size (Length) Associated with 80% POD for Each Vendor. Bounds are 95%.

Case	Lower Bound, mm (in.)	Flaw Size, mm (in.)	Upper Bound ($a_{90/97.5}$), mm (in.)
ARLW	20.0 (0.79)	23.4 (0.92)	29.0 (1.14)
DCSI	21.0 (0.83)	25.5 (1.00)	33.1 (1.30)
NBIE	15.3 (0.60)	18.0 (0.71)	22.2 (0.87)
TUQZ	19.7 (0.78)	23.8 (0.94)	30.7 (1.21)
YPJH	20.5 (0.81)	24.6 (0.97)	31.4 (1.23)
All	20.0 (0.79)	23.1 (0.91)	27.5 (1.08)

F.3 Effect of Explanatory Variables on POD

In this section, the effect of the most important explanatory variables in the experiment are evaluated. These explanatory variables are team, flaw orientation, flaw location, and flaw "OnEdge" (i.e., is the flaw on the edge of the weld). The evaluation is presented with categorical

tables that display POD and a standard deviation of the estimate. These tables do not consider flaw size, and might therefore lead to incorrect conclusions if the flaw sizes of the different categories of flaws differ greatly. These tables are meant to provide an overview of the effect of these variables.

Table F-9 presents POD by team and flaw orientation. Except for team NBIE, circumferential POD is greater than axial, but not by much. In fact, if all teams results are combined (presented in the "All Teams" row of the table), we see that axial and circumferential POD differ by only 3 percentage points, not a significant amount.

Table F-9 POD of Axial/Circumferential Flaws by Team

	A	C
ARLW	65±11	75±6
DCSI	70±10	73±6
NBIE	95±6	82±5
TUQZ	75±10	82±5
YPJH	70±10	80±5
All Teams	75±4	78±2
NOBS	20	55

Table F-10 presents PODs for flaw locations. From the "All Teams" row, there is weak evidence that flaws "In Surface Features" are harder to detect than flaws at the other location. There is no evidence that ground welds behave any differently than unground welds, or that detection in the HAZ is different than these other two locations. Tables F-9 and F-10 are comparable to tables presented in Phase II and can therefore be used to compare Phase II performance with Phase III.

Tables F-9 and F-10 do not account for any relationships that might exist between flaw orientation and locations. For example, all axial flaws are in welds. To better understand the effect that orientation and location jointly have on POD, Table F-11 was produced. From the "All Teams" row in Table F-11, we would conclude that POD for axial flaws are hardest to detect when in a surface feature (POD is reduced from 80% to 50%). For circumferential flaws, HAZ and surface feature locations produce about the same POD (approximately 70%), while flaws in the weld have a POD of 80%. To compare axial with circumferential, we might use the NGWeld location, and for this location axial has a slightly higher POD (86% vs. 80%). This is not a significant difference.

Table F-12 presents the effect of weld edge on POD. "Weld Edge" is actually a categorization of flaw location. Circumferential flaws on the weld edge are thought to be difficult to identify. As in Phase II, we see that this indeed is the case. The POD difference appears to be significant at 5% level. Note that the teams' POD show a consistent trend: POD(Edge) < POD(NotEdge).

Table F-10 POD of Flaws in Different Locations: in Ground Weld, in Unground Weld, in HAZ, in Surface Feature

	Ground Weld	HAZ	In Surface Feature	Unground Weld
ARLW	80±19	75±7	65±10	73±14
DCSI	80±19	72±8	65±10	82±12
NBIE	100±12	83±6	78±9	100±6
TUQZ	80±19	83±6	70±10	91±10
YPJH	80±19	81±7	70±10	82±12
All Teams	84±8	79±3	70±4	85±5
NOBS	5	36	23	11

Table F-11 POD of Flaws in Different Orientations/Locations

	A: Ground Weld	A: In Surface Feature	A: Unground Weld	C: HAZ	C: In Surface Feature	C: Unground Weld
ARLW	80±19	40±22	70±15	75±7	72±11	100±43
DCSI	80±19	40±22	80±13	72±8	72±11	100±43
NBIE	100±12	80±19	100±7	83±6	78±10	100±43
TUQZ	80±19	40±22	90±11	83±6	78±10	100±43
YPJH	80±19	40±22	80±13	81±7	78±10	100±43
All Teams	84±8	48±10	84±5	79±3	76±5	100±12
NOBS	5	5	10	36	18	1

Table F-12 POD of Circumferential Flaws on Weld Edge

	Not on Edge	On Edge
ARLW	79±6	58±14
DCSI	79±6	50±14
NBIE	84±6	75±13
TUQZ	88±5	58±14
YPJH	86±5	58±14
All Teams	83±3	69±6
NOBS	43	12

F.4 Location Errors

This section examines the y/x (i.e., circumferential/axial) location error. To define location error, we calculated flaw/indication midpoints and subtracted the two (i.e., indication − flaw midpoint). Figure F-7 provides plots of this location error, separated by axial/circumferential flaw orientation. As one can see from the plots, the error in the circumferential (Y) direction has a few “gross” errors around 20 mm (0.8 in.). If these gross errors were eliminated, the remaining errors would

be bounded by ±10 mm (0.39 in.). Roughly the same statement could be made for the transverse (X) direction, except with a threshold of 12 mm (0.47 in.) to define the gross-error threshold.

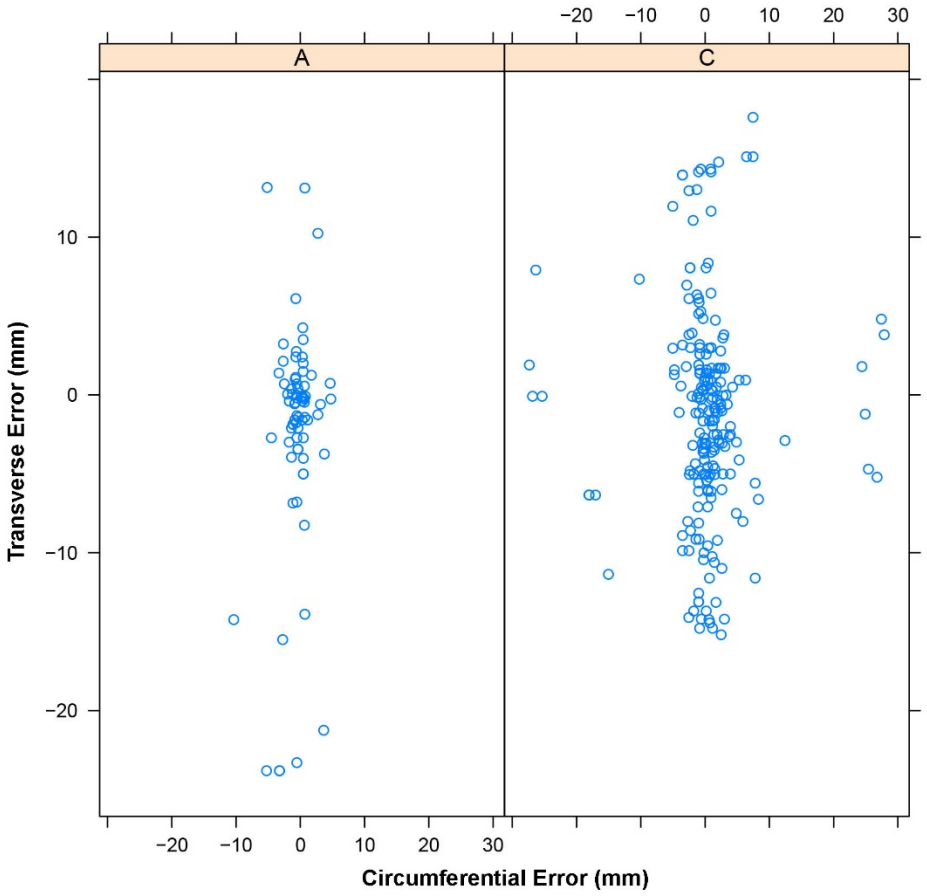


Figure F-7 Plot of Location Errors of Axial and Circumferential Flaws

Realize that the error distributions being presented in this section are truncated by the scoring procedure used for detection; indication and flaw have to be within a certain distance of each other to be associated with each other and allow a location error to be computed. Very large location errors will result in a lowering of POD and not affect the location error distribution being calculated.

Tables F-13 and F-14 present statistics of the location error distribution. Table F-13 presents root mean squared error (RMSE) of the location error values. Because there is little bias in the location error, one can think of RMSE as a standard deviation with units of mm or inches. The location errors in both tables are separated by flaw orientation (Axial and Circumferential).

The “All Teams” row in Table F-13 provides the best overall summary of location errors. For axial flaws, the RMSE in location error is seen to be 2.25 mm (0.09 in.) in the Y dimension and 7.52 mm (0.30 in.) in the X direction, while for circumferential flaws it is 6.66 and 6.76 mm (0.26 and 0.27 in.), respectively. From these values, one would conclude that there is a greater problem with location error in the X-direction than Y on axial cracks, but not on circumferential cracks. If we look at individual teams, it appears that TUQZ has the most problems with location error in the X-direction; TUQZ’s X location error is twice the Y location error.

Table F-14 presents the quantiles of the absolute value of the location error to quantify the distribution shape. From the quantile values other than 100%, one can see that the location error in the X-direction is larger than that in the Y-direction by approximately a factor of two.

Table F-13 RMSE Location Error for Axial and Circumferential Flaws

	Circumferential (Y) Dimension		Transverse (X) Dimension	
	Axial	Circumferential	Axial	Circumferential
ARLW	3.14	7.03	5.41	5.63
DCSI	3.05	6.98	9.43	8.24
NBIE	1.73	7.31	7.71	5.17
TUQZ	1.67	5.13	6.57	6.73
YPJH	1.29	6.70	7.76	6.73
All Teams	2.25	6.66	7.52	6.76

Table F-14 Quantiles of Absolute (Location Error) for Axial and Circumferential Flaws

	Circumferential (Y) Dimension		Transverse (X) Dimension	
	Axial	Circumferential	Axial	Circumferential
0%	0.15	0.00	0.00	0.00
25%	0.50	0.80	0.57	1.60
50%	0.70	1.45	1.85	3.60
75%	1.88	2.80	3.98	7.43
100%	10.35	27.90	23.80	17.60

F.5 Length Sizing Error

In this section, an overview of length sizing error is provided for visual testing. Figure F-8 plots measured length vs. true length for all detected flaws. The red line in the plot represents a perfect measurement (true=measured), while the black line represents a linear regression fit to the data (meas = $\beta_1 + \beta_2 * \text{true}$). As one can see, from the regression fit, inspectors tend to oversize small flaws and undersize large flaws, an almost universal characteristic for NDE flaw sizing.

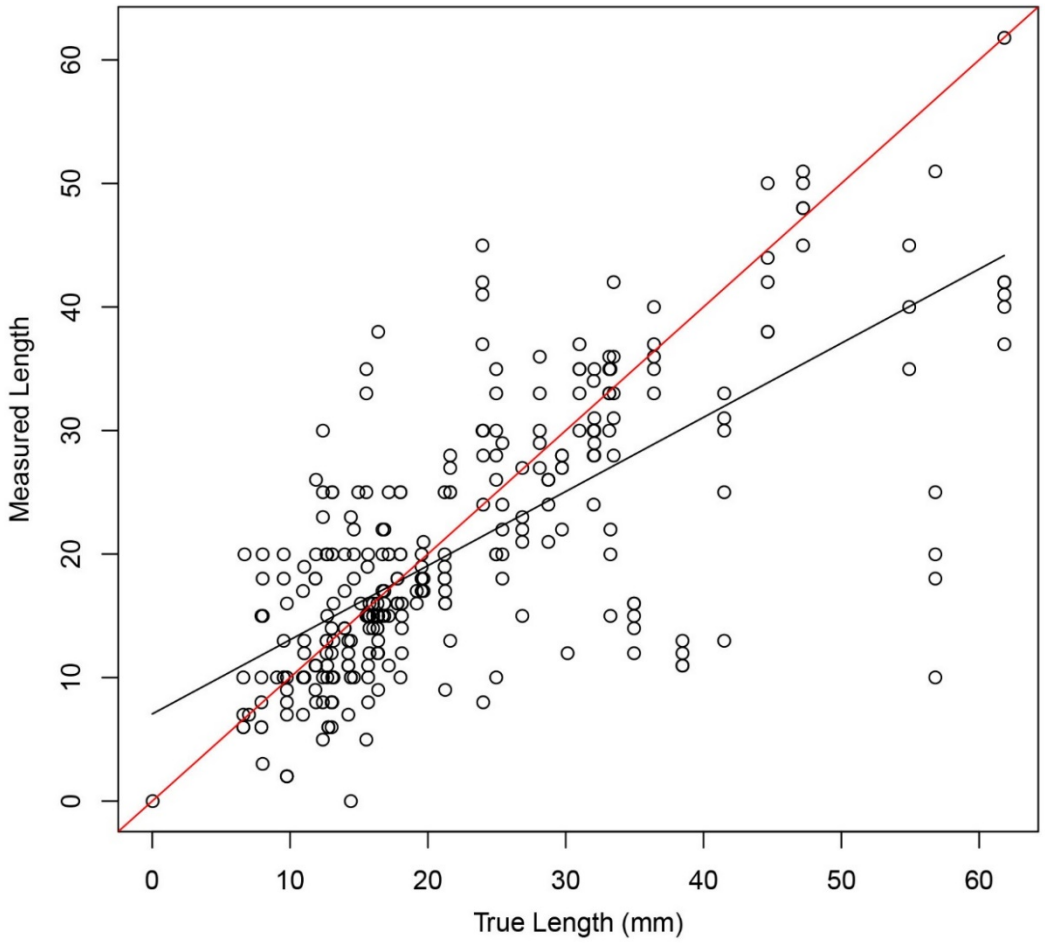


Figure F-8 Measured Lengths vs. True Lengths. Black line is regression fit; red line is measured=true.

Tables F-15 and F-16 present a few important sizing error statistics. Sizing error is defined as *measured length – true length*. Bias is mean of the error, while RMSE is defined as

$$\mathrm{RMSE}(\mathrm{error}) = \sqrt{\mathrm{mean}(\mathrm{error})^2} \qquad \text{(F.4)}$$

Table F-15 Sizing Error Statistics by Team

	NOBS	Bias, mm (in.)	Standard Deviation, mm (in.)	RMSE, mm (in.)
ARLW	55	0.2 (0.01)	10.4 (0.41)	10.3 (0.41)
DCSI	55	−2.6 (−0.10)	9.8 (0.39)	10.0 (0.39)
NBIE	65	−2.9 (−0.11)	7.8 (0.31)	8.2 (0.32)
TUQZ	62	−2.8 (−0.11)	8.8 (0.35)	9.1 (0.36)
YPJH	61	−1.3 (−0.05)	8.9 (0.35)	8.9 (0.35)

Table F-16 Sizing Error Statistics by Orientation

	NOBS	Bias, mm (in.)	Standard Deviation, mm (in.)	RMSE, mm (in.)	Average Length, mm (in.)
Axial	76	2.4 (0.09)	7.3 (0.29)	7.6 (0.30)	13.5 (0.53)
Circumferential	222	−3.4 (−0.13)	9.1 (0.36)	9.8 (0.39)	25.5 (1.00)

Perhaps the main conclusion to draw from Table F-15 is that team sizing performances are similar to each other, with an RMSE of about 9 mm (0.35 in.). In fact the differences displayed in this table would not be significant at a 95% confidence level. It appears that NBIE produced the best performance with an RMSE of 8.2 mm (0.32 in.).

Table F-16 examines the difference between axial and circumferential sizing capability. Sizing of axial flaws achieves a smaller RMSE than circumferential, but axial flaws are smaller than circumferential, so this difference may be due to flaw size instead of orientation. Also, axial flaws are with the weld so inspectors have a limit on flaw size (i.e., width of weld) that they can use in their sizing decisions. The last column in the table presents the mean (true) size of axial/circumferential flaws. As one can see, axial flaws are on average about half the size of circumferential flaws. If RMSE were expressed as a relative error (RMSE/mean), the axial RMSE would then be greater than circumferential.

Table F-17 summarizes the average time taken per test specimen in Phase III by each team as well as the minimum and maximum inspection times. Comparing the data in this table to that presented in Tables F-3, F-5, and F-7, it is obvious that there is not a strong correlation between average time spent on a specimen and the performance. Instead, the variation in inspection times is likely a function of several factors, including the time taken to evaluate an indication and the time taken to record the information on the data sheet.

Table F-17 Minimum, Maximum, and Average Inspection Time per Specimen for Primary Analyst

	Time Per Specimen (min)		
Teams	Average	Minimum	Maximum
DCSI	17	6	36
YPJH	24	12	44
ARLW	16	8	31
NBIE	30	8	62
TUQZ	27	10	67
Total	23	6	67

APPENDIX G EVALUATION OF LOGISTIC REGRESSION MODELS

This appendix describes attempts to find a better logistic regression model for describing the relationship between POD and flaw size (COD or Length) in this study. A good regression model should fit the present data, be mathematically simple, not violate physical principles, and ideally be "related to" the models we have used in previous sections.

The motivation for this investigation is the somewhat poor fit exhibited by the logistic regression model for length:

$$POD(Len) = \text{logistic}(\beta_1 + \beta_2 \times Len) \tag{G.1}$$

where Len is the length variable, and β_0 and β_1 are the model parameters. This model can provide unrealistically high probabilities for long cracks. For example, compare the POD curve vs. the data in Figure G-1. One of the longest cracks (55 mm [2.165 in.]) has an empirical POD of 60%, but the POD model assigns a probability of 99.99%, not a realistic fit. A log model of the form

$$POD(Len) = \log istic(\beta_1 + \beta_2 \times \log(Len)) \tag{G.2}$$

provides a much more reasonable fit to large cracks, as Figure G-2 illustrates. Unfortunately, a log model cannot fit the data near Len = 0. The log model forces POD(Len = 0)) to be zero, which is physically impossible. So one would like to identify a logistic regression model that better fits the data at both ends of the size range.

G.1 Improved Logistic Regression Models

G.1.1 Overview of Model Fits

To identify the best model for the data, the fits of 13 prospective models using Phase III data only are presented in Table G-1. How well the model fits the data can be assessed by the dispersion and AIC (Akiake Information Criteria) statistics. In this discussion, we will concentrate on the dispersion statistic to evaluate goodness of fit. Dispersion measures the scatter of data about the regression curve, as a residual mean square statistic does for regular regression. Roughly speaking, if the model fits the data well, the dispersion should be around 1; if it is large (above say, 1.5 for the degrees of freedom [DOF] in this data) the model does not fit the data. The model may not fit the data because 1) the regression curve does not fit, or 2) the binomial distribution does not fit the data.

The data plots in Figure G-1 show that non-binomial variability exists in the data. Each data point in this figure represents replicate measurements on a single flaw. From the figure we can see that flaws of essentially the same length can exhibit dramatically different PODs (see the flaws around 30 mm and 55 mm [1.2 in. and 2.165 in.]). This flaw-to-flaw variability exists because flaw length is not a particularly good predictor of POD. Consequently, a model using only the variable flaw length cannot hope to achieve a dispersion near the theoretical expected value of 1; a value near 2 seems to indicate a good fit when binomial variability is accounted for.

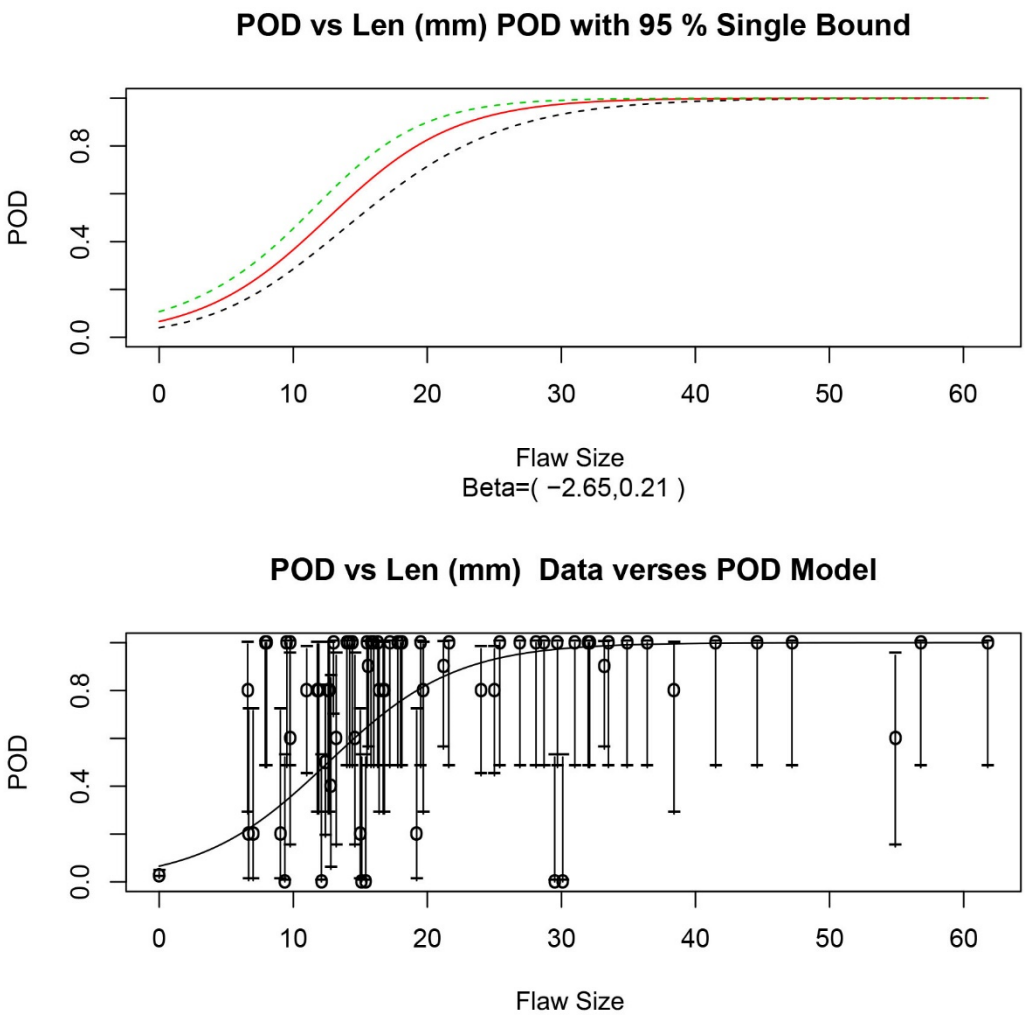


Figure G-1 POD vs. Length using Linear Model

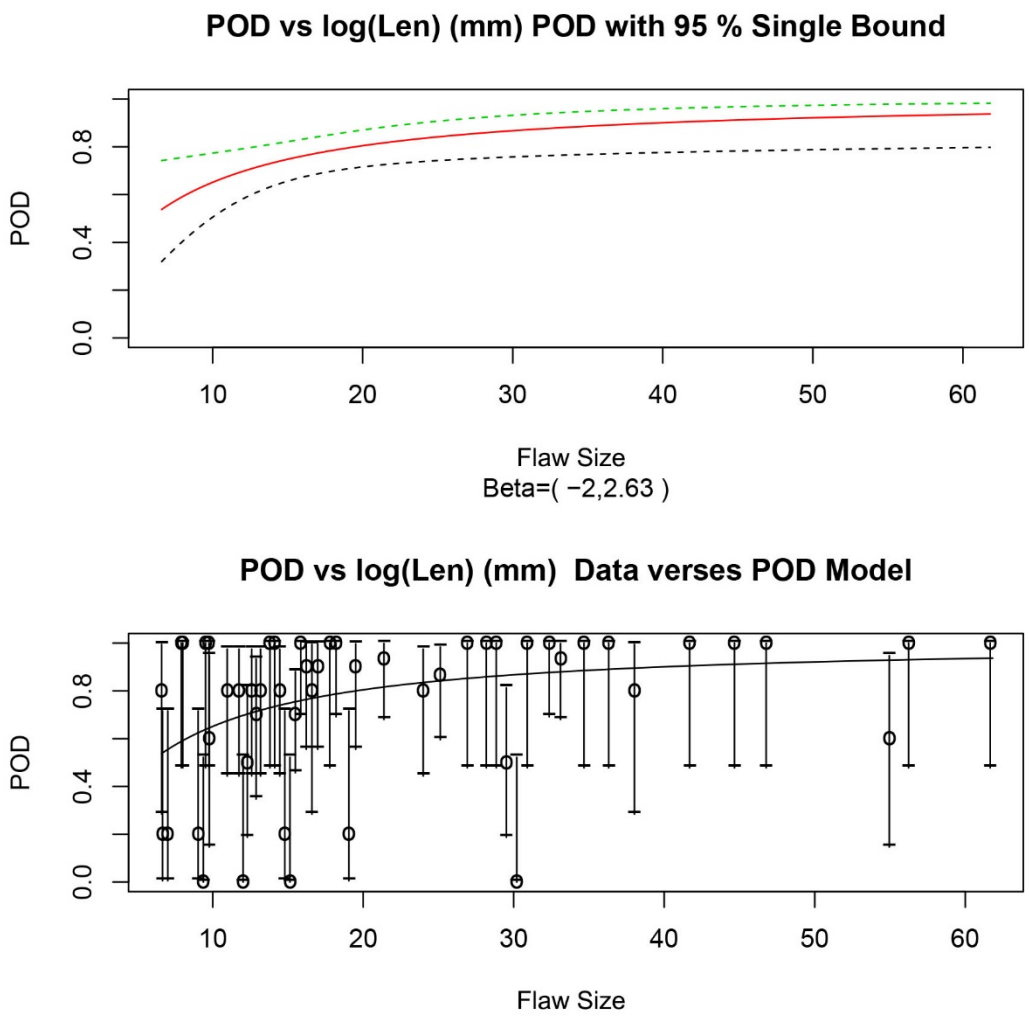


Figure G-2 POD vs. Length Using Log Model

Table G-1 Summary of Various POD Model Fits to Data

	Model	DOF	Dispersion	AIC
1	POD = logistic $(\beta_1 + \beta_2\, COD)$	116	1.97	314.45
2	POD = logistic $(\beta_1 + \beta_2\, Len)$	116	3.23	460.57
3	POD = logistic $(\beta_1 + \beta_2\, Area)$	116	2.49	374.36
4	POD = logistic $(\beta_1 + \beta_2\, COD + \beta_3\, Len)$	69	4.03	361.55
5	POD = $\in + (1 - \in)\, \text{log}istic(\beta_1 + \beta_2 \log(COD))$	73	3.11	283.62
6	Box-Cox Model Fits:			
7	POD = logistic $\left(\beta_1 + \beta_2\, COD^{\beta_3}\right)$	115	1.84	299.20
8	POD = logistic $\left(\beta_1 + \beta_2\, Len^{\beta_3}\right)$	115	2.32	354.33
9	POD = logistic $\left(\beta_1 + \beta_2\, Area^{\beta_3}\right)$	115	1.86	302.21
10	Fits without False Calls:			
11	POD = logistic $(\beta_1 + \beta_2 \log(COD))$	73	2.41	232.94
12	POD = logistic $(\beta_1 + \beta_2\, COD)$	73	2.33	226.54
13	POD = logistic $(\beta_1 + \beta_2 \log(Len))$	73	3.11	283.61
14	POD = logistic $(\beta_1 + \beta_2\, Len)$	73	3.14	286.14
15	Model 2 Dispersion for Len > 0 Points	73	4.18	

It should be noted that a few other obvious models not listed in Table G-1 were also considered. Logistic models with a quadratic polynomial were also fitted, but they produced curves that were not monotonically increasing, and did not have an improved dispersion. We also considered spline models. Spline models can be made to fit the data well, but use of spline models requires an arbitrary assignment of spline knots.

From Table G-1, one can see that the "Box-Cox" model fits the data best, regardless of the choice of flaw size variable (COD, Length, or Area). The Box-Cox model is based on the Box-Cox transformation used in statistics to "stabilize variance." In its application in the logistic regression model, this produces a model that can closely resemble a log function when the power coefficient, β_3, is small. See Figure G-3 (Model 13 vs. Model 8) as an example of this. Thus one can view the Box-Cox model as a more general case of both the logistic-linear and the logistic-log models.

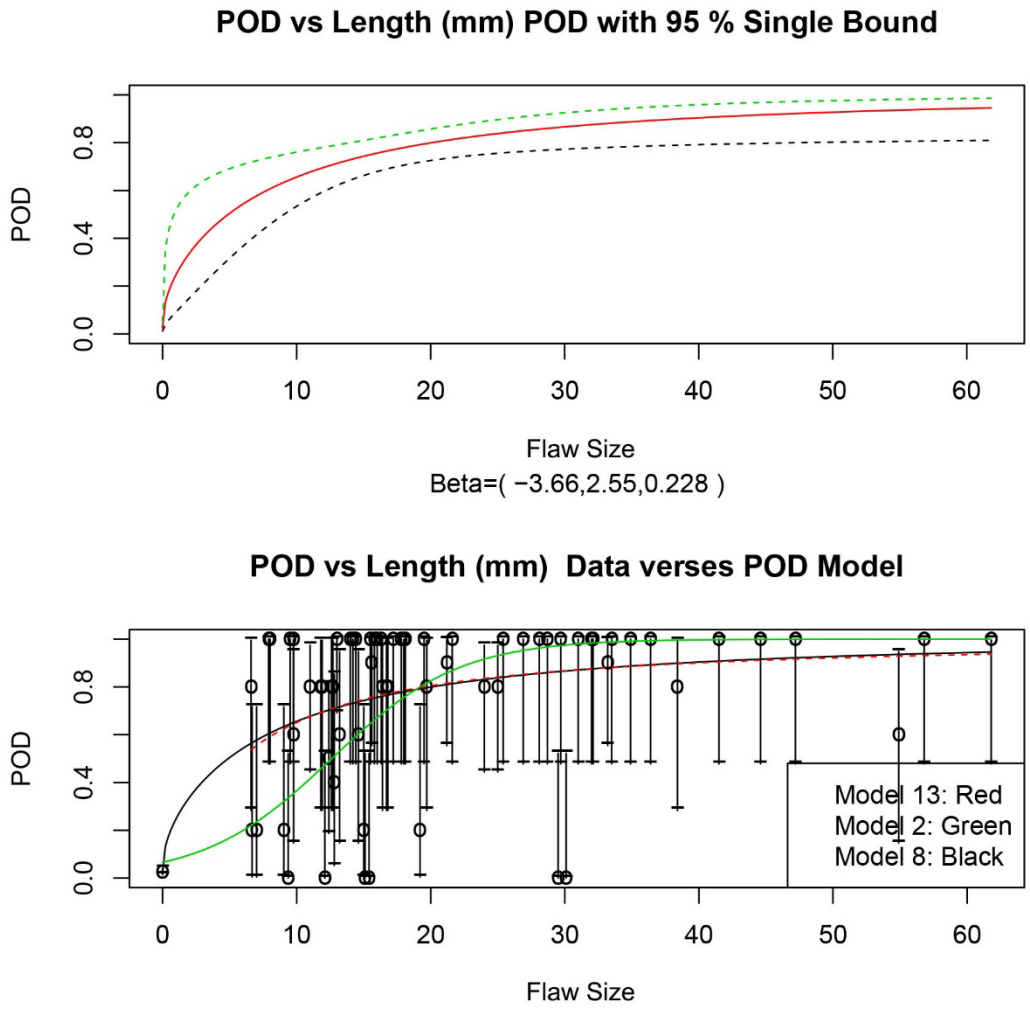


Figure G-3 POD vs. Flaw Length with Box-Cox Model

For COD, the Box-Cox model (7) attains a dispersion of 1.84, slightly better than value of 1.97 attained by the linear (1) model we have used in this report. The Box-Cox model dispersion for (7) is much better than that exhibited for the log model (11).

For Length, the Box-Cox model (8) has a dispersion of 2.32, which is much better than either the linear model (2) or the log (13). It is clear that the Box-Cox model is a better fit than the other models involving length. One should mention model (5) presented in Table G-1. This "offset" model is one way to modify the log-model in 13 so that it accounts for non-zero false calls. (The offset, ϵ, represents false call probability.) For these data, the offset set model fits as well as the log model and would also represent a methodology for creating a model that behaves properly at zero. However, it does not fit as well as model 8.

If the false call data are eliminated from the data set, the fits presented in lines 11–14 of Table G-1 are produced. It is interesting to compare model 13 (the log-length model) to model 14 (which is model 3 without false call data). Note that both fit the data equally well, indicating that a linear model can fit the data as well as the log over a similar size range.

G.1.2 Box-Cox Model Fits

From Table G-1 we see that the three-parameter Box-Cox model should be used to describe the relationship between POD and flaw size. This section provides plots of these fits to Length and COD as illustrated in Figures F-3 and G-4. To allow the reader to gauge the changes this would make to the POD curve, the Box-Cox fits are plotted along with the standard regression model curve fits. Thus for COD, model 7 is plotted against model 1, while for Length, model 8 is plotted against model 2.

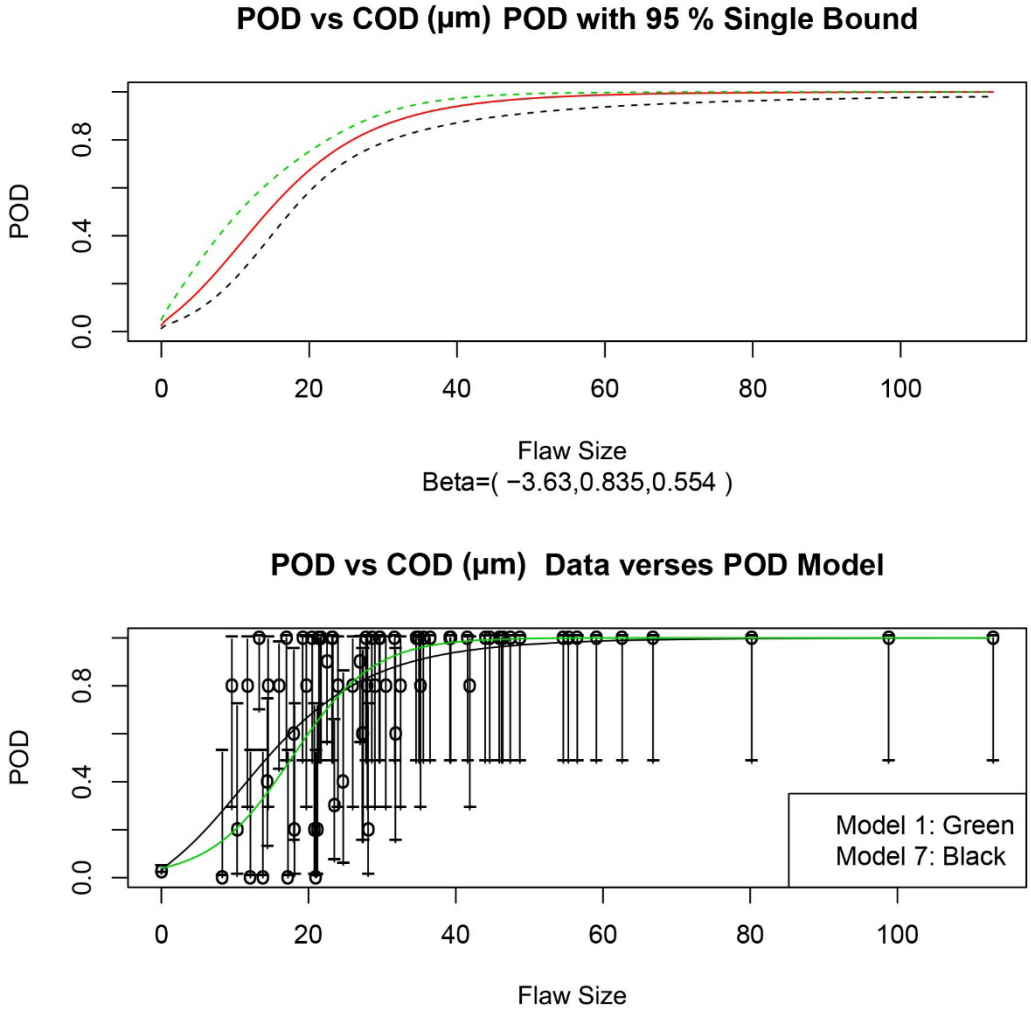


Figure G-4 POD vs. COD Using Box-Cox Model

Individual fits of the Box-Cox model are also performed to each team's data. These results are presented in Figures G-5 and G-6. The flaw sizes associated with an 80% POD are presented in Tables G-2 and G-3 for these fits. These can be compared with the data presented in Tables F-5 and F-7, which are based on the regression fits using Models 1 and 2.

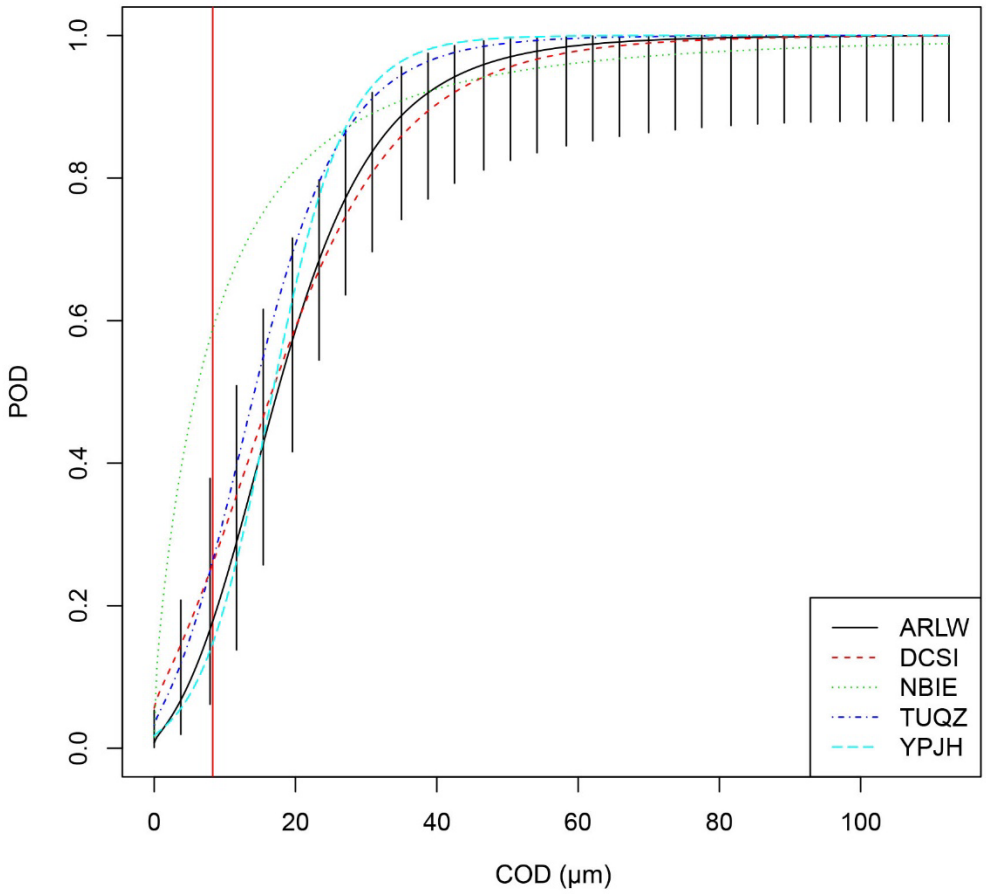


Figure G-5 POD vs. COD for Each Team

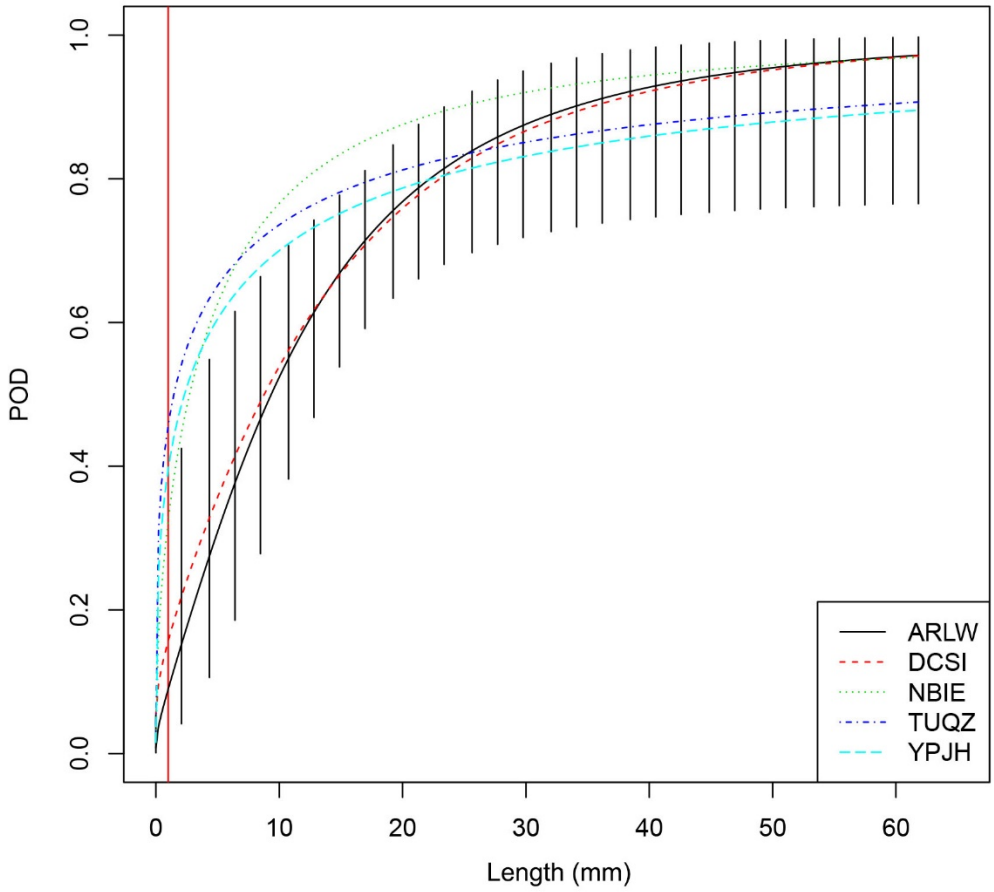


Figure G-6 POD vs. Length for Each Team

In the data presented in Figure G-5 and G-6, the red vertical line identifies the size of the smallest flaw present in the data. The hatched lines about team ARLW represent 95% confidence bounds for that POD curve. These 95% bounds provide the reader with a rough idea of fit uncertainty. From the figures, we see that the teams break into possibly two groups (ARLW,DCSI) vs. (NBIE,TUQZ,YPJH) based on their POD curves.

Table G-2 Estimate of Crack Size (COD, Microns*) Associated with 80% POD for Each Team. Bounds are 95%.

Case	Lower Bound	Flaw Size	Upper Bound
ARLW	23.5	28.6	44.0
DCSI	24.2	30.3	67.4
NBIE	4.9	19.0	46.3
TUQZ	18.7	23.6	44.4
YPJH	20.5	24.2	70.6
All	22.5	25.9	31.0

*To convert microns to inches, multiply microns by 0.00004.

Table G-3 Estimate of Crack Size (Length) Associated with 80% POD for Each Team. Bounds are 95%.

Case	Lower Bound, mm (in.)	Flaw Size, mm (in.)	Upper Bound, mm (in.)
ARLW	16.2 (0.64)	22.2 (0.87)	61.8 (2.43)
DCSI	16.3 (0.64)	23.0 (0.91)	43.8 (1.73)
NBIE	0.7 (0.03)	12.1 (0.48)	26.5 (1.04)
TUQZ	10.3 (0.41)	17.7 (0.70)	24.0 (0.94)
YPJH	9.9 (0.39)	22.4 (0.88)	25.0 (0.99)
All	14.0 (0.55)	20.1 (0.79)	47.6 (1.87)

G.1.3 Length Compared to COD Regressions

From Table G-1, one can see that the COD size variable describes POD much better than Flaw Length (compare the dispersion of model 7 to that of model 8). Thus the "poor" fit of the length model is not due to any deficiency in the regression model form, but due to the fact that detectability is more closely related to COD than flaw length. Whenever possible, it would be better to use a POD curve involving COD instead of one involving flaw length. For example, if one required a POD(Length) curve for safety calculations, it might be best to derive it from the POD(COD) curve, particularly if the (COD,Length) distribution relevant to the safety calculation was different than that occurring in this study. The relevant formula would be:

$$\mathrm{POD}(\mathrm{Len}) = \int \mathrm{POD}(\mathrm{COD}) \int (\mathrm{COD} \mid \mathrm{Len})\, d\mathrm{COD} \qquad \text{(G.3)}$$

where ∫(COD|Len) is the conditional distribution of COD given Length.

G.2 Summary

The results presented in this appendix indicate that the Box-Cox model fits may best represent the variability in the data from Phase III, and allow the regression fits for the POD curves to better represent the POD at both the high and low ends of the flaw variable (COD or length). Comparisons of the estimated flaw size associated with 80% POD at a 95% confidence level indicate that this value does not change significantly, regardless of whether the linear models or

the Box-Cox models are used. This finding appears to indicate that the true value of the Box-Cox models is in capturing the (non-binomial) variability associated with the Phase III data (and possibly the Phase II data as well) to provide a more accurate estimate of the POD at the high and low ends of the flaw parameters.

This additional analysis also appears to indicate that length may not be a good predictor of detectability in RVT, as the POD appears to be more closely related to COD and perhaps, area. This result seems to hold regardless of the model type used (linear, logarithmic, or Box-Cox), and assessing the POD as a function of length (for safety calculations, or other purposes) may require assessing the relationship between COD and length first.

NRC FORM 335
(12-2010)
NRCMD 3.7

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

(See instructions on the reverse)

1. REPORT NUMBER
(Assigned by NRC, Add Vol., Supp., Rev., and Addendum Numbers, if any.)

NUREG/CR-7246

2. TITLE AND SUBTITLE

Reliability Assessment of Remote Visual Examination

3. DATE REPORT PUBLISHED

MONTH	YEAR
August	2018

4. FIN OR GRANT NUMBER

5. AUTHOR(S)

P. Ramuhalli, P. G. Heasler, T. L. Moran, A. E. Holmes, M. T. Anderson

6. TYPE OF REPORT

Technical

7. PERIOD COVERED (Inclusive Dates)

8. PERFORMING ORGANIZATION - NAME AND ADDRESS (If NRC, provide Division, Office or Region, U. S. Nuclear Regulatory Commission, and mailing address; if contractor, provide name and mailing address.)

Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352

9. SPONSORING ORGANIZATION - NAME AND ADDRESS (If NRC, type "Same as above", if contractor, provide NRC Division, Office or Region, U. S. Nuclear Regulatory Commission, and mailing address.)

Division of Engineering
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555-0001

10. SUPPLEMENTARY NOTES

11. ABSTRACT (200 words or less)

Remote visual testing (RVT) is a commonly used nondestructive examination method for inservice inspection (ISI) of reactor internals to detect cracking and gross component failures. Despite widespread use, the detection reliability of RVT and the factors that impact overall RVT performance have been unresolved issues. This report describes the results from an assessment sponsored by the U.S. Nuclear Regulatory Commission and conducted by the Pacific Northwest National Laboratory, in cooperation with the Electric Power Research Institute, for evaluating the reliability of RVT methods currently being used for [reactor] in-vessel visual inspection.

This report describes the design and noted limitations of the three phases of research, the analysis methodology for each phase, and the results of the research. The results of this assessment provide a benchmark set of data on the reliability of RVT for detecting cracking, assuming the implementation of field-like inspection procedures. The likely impact of several uncontrolled factors on RVT detection performance are discussed, and recommendations regarding the use of these results to assess field performance are provided. Finally, recent advances in RVT technology are briefly discussed and point to the potential need for continued research to evaluate the capability and effectiveness of the technique as improvements are implemented.

12. KEY WORDS/DESCRIPTORS (List words or phrases that will assist researchers in locating the report.)

Nondestructive Examination, Remote Visual Testing, Inservice Inspection, In-vessel visual inspection, Probability of Detection, Crack Opening Displacement

13. AVAILABILITY STATEMENT

unlimited

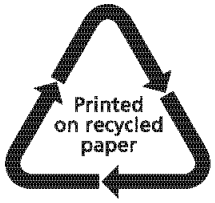
14. SECURITY CLASSIFICATION

(This Page)
unclassified

(This Report)
unclassified

15. NUMBER OF PAGES

16. PRICE



Federal Recycling Program

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, DC 20555-0001

OFFICIAL BUSINESS