Statistically Based Reevaluation of PISC-II Round Robin Test Data

Prepared by P. G. Heasler, T. T. Taylor, S. R. Doctor

Pacific Northwest Laboratory
Operated by
Battelle Memorial Institute

Prepared for
U.S. Nuclear Regulatory Commission

AVAILABILITY NOTICE

Availability of Reference Materials Cited in NRC Publications

Most documents cited in NRC publications will be available from one of the following sources:

1. The NRC Public Document Room, 2120 L Street, NW., Lower Level, Washington, DC 20555
2. The Superintendent of Documents, U.S. Government Printing Office, P.O. Box 37082, Washington, DC 20013-7082
3. The National Technical Information Service, Springfield, VA 22161

Although the listing that follows represents the majority of documents cited in NRC publications, it is not intended to be exhaustive.

Referenced documents available for inspection and copying for a fee from the NRC Public Document Room include NRC correspondence and internal NRC memoranda; NRC bulletins, circulars, information notices, inspection and investigation notices; licensee event reports; vendor reports and correspondence; Commission papers; and applicant and licensee documents and correspondence.

The following documents in the NUREG series are available for purchase from the GPO Sales Program: formal NRC staff and contractor reports, NRC-sponsored conference proceedings, international agreement reports, grant publications, and NRC booklets and brochures. Also available are regulatory guides, NRC regulations in the *Code of Federal Regulations*, and *Nuclear Regulatory Commission Issuances*.

Documents available from the National Technical Information Service include NUREG-series reports and technical reports prepared by other Federal agencies and reports prepared by the Atomic Energy Commission, forerunner agency to the Nuclear Regulatory Commission.

Documents available from public and special technical libraries include all open literature items, such as books, journal articles, and transactions. *Federal Register* notices, Federal and State legislation, and congressional reports can usually be obtained from these libraries.

Documents such as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings are available for purchase from the organization sponsoring the publication cited.

Single copies of NRC draft reports are available free, to the extent of supply, upon written request to the Office of Administration, Distribution and Mail Services Section, U.S. Nuclear Regulatory Commission, Washington, DC 20555.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at the NRC Library, 7920 Norfolk Avenue, Bethesda, Maryland, for 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 the American National Standards Institute, 1430 Broadway, New York, NY 10018.

DISCLAIMER NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability of responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

Statistically Based Reevaluation of PISC-II Round Robin Test Data

Manuscript Completed: March 1993
Date Published: May 1993

Prepared by
P. G. Heasler, T. T. Taylor, S. R. Doctor

Pacific Northwest Laboratory
Richland, WA 99352

Prepared for
Division of Engineering
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555
NRC FIN B2289

Abstract

This report presents a re-analysis of an international PISC-II round-robin inspection results using formal statistical techniques to account for experimental error. The analysis examines:

- U.S. team performance vs. other participants performance
- flaw sizing performance and errors associated with flaw sizing
- factors influencing flaw detection probability
- performance of all participants with respect to recently adopted ASME Section XI flaw detection performance demonstration requirements

and develops conclusions concerning ultrasonic inspection capability.

Contents

Abstract	iii
Executive Summary	xi
Acknowledgements	xiii
1.0 Introduction	1
2.0 Contents of the PISC-II Data Set	3
2.1 Variables Recorded in the PISC-II Inspection Data Sets	3
2.1.1 Inspection Variables	4
2.2 Material Variables: Description of PISC-II Blocks	6
2.3 Scoring: Calculation of Inspection Performance Statistics	9
2.3. Scoring with Multiple Flaws	11
2.4 Comparison of PNL-Scored Results to PISC-II Results	11
3.0 Analysis of Detection Performance	15
3.1 POD Regression Methodology	15
3.2 Comparison of U.S. with Non-U.S. Teams	16
3.3 Comparison of CLAD POD with Base-Metal POD	21
3.4 Team-to-Team Variability	23
3.5 What Procedures were the Best?	23
3.6 Summary	24
4.0 Sizing Performance	37
4.1 Regression Models and Statistics for Evaluating Sizing	37
4.2 Comparison of U.S. with Non-U.S. Sizing Capability	39
4.3 Comparison of Clad Versus Base-Metal Sizing Capability	42
4.4 Comparison of Sizing Procedures	42
4.5 Summary	44
5.0 Use of ASME Criteria to Evaluate PISC-II Procedures	53
5.1 ASME Performance Demonstration Results for Clad/Base Metal Interface Region	55
5.1.1 Detection Results for the Clad/Base Metal Region	55
5.1.2 Sizing Results for the Clad/Base Metal Region	57
5.2 Detection Results for ASME-Type Vessel Shell Weld Test	57
5.2.1 ASME Vessel Shell Sizing Results	57
5.3 Summary	57
6.0 Conclusions and Recommendations	61
6.1 Conclusions Regarding Probability of Detection	61
6.2 Conclusions Regarding Sizing	61
6.3 Conclusions Regarding PISC-II Data and ASME Section XI	61
6.4 Conclusions Regarding Experimental Design	62
6.5 Conclusions Regarding Practice and Recommendations for Regulation of Ultrasonic Inspection of Reactor Pressure Vessels	62

7.0 References	65
Appendix A. Flaw and Grading Unit Plots	A.1
Appendix B. POD Curve Fits	B.1
Appendix C. Statistical Methods	C.1
Appendix D. Sizing Curve Fits	D.1
Appendix E. ASME Section XI Code Performance Demonstration Requirements for Clad/Base-Metal Region and Pressure Vessel Shell Welds	E.1

Figures

2.1. Scoring with Multiple Indications	10
2.2. Scoring with High False Calls	10
2.4. Plot of Flaw Lengths for PNL Versus PISC-II Scored Data	13
2.5. Plot of Flaw Depths for PNL Versus PISC-II Scored Data	13
3.1. U.S. POD, Base Material, Rough Crack, Manual-20% DAC Procedure	28
3.2. Non-U.S. POD, Base Material, Rough Crack, Manual-20% DAC Procedure	28
3.3. U.S. POD, Base Material, Smooth Crack, Manual-20% DAC Procedure	29
3.4. Non-U.S. POD, Base Material, Smooth Crack, Manual-20% DAC Procedure	29
3.5. U.S. POD, Base Material, Volumetric Crack, Manual-20% DAC Procedure	29
3.6. Non-U.S. POD, Base Material, Volumetric Crack, Manual-20% DAC Procedure	29
3.7. U.S. POD, Clad Material, Rough Crack, Manual-20% DAC Procedure	30
3.8. Non-U.S. POD, Clad Material, Rough Crack, Manual-20% DAC Procedure	30
3.9. U.S. POD, Clad Material, Smooth Crack, Manual-20% DAC Procedure	30
3.10. Non-U.S. POD, Clad Material, Smooth Crack, Manual-20% DAC Procedure	30
3.11. U.S. POD, Clad Material, Volumetric Crack, Manual-20% DAC Procedure	31
3.12. Non-U.S. POD, Clad Material, Volumetric Crack, Manual-20% DAC Procedure	31
3.13. U.S. POD, Base Material, Rough Crack, Advanced Procedure	31
3.14. Non-U.S. POD, Base Material, Rough Crack, Advanced Procedure	31
3.15. U.S. POD, Base Material, Smooth Crack, Advanced Procedure	32
3.16. Non-U.S. POD, Base Material, Smooth Crack, Advanced Procedure	32
3.17. U.S. POD, Base Material, Volumetric Crack, Advanced Procedure	32
3.18. Non-U.S. POD, Base Material, Volumetric Crack, Advanced Procedure	32
3.19. U.S. POD, Clad Material, Rough Crack, Advanced Procedure	33
3.20. Non-U.S. POD, Clad Material, Rough Crack, Advanced Procedure	33
3.21. U.S. POD, Clad Material, Smooth Crack, Advanced Procedure	33

3.22. Non-U.S. POD, Clad Material, Smooth Crack, Advanced Procedure	33
3.23. U.S. POD, Clad Material, Volumetric Crack, Advanced Procedure	34
3.24. Non-U.S. POD, Clad Material, Volumetric Crack, Advanced Procedure	34
3.25. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Rough Flaws	34
3.26. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Smooth Flaws	34
3.27. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Volumetric Flaws	35
3.28. Comparison of Procedures for Rough Flaws in Base Metal	35
3.29. Comparison of Procedures for Smooth Flaws in Base Metal	35
4.1. Standard Deviation of Measured Depth Versus True Depth	38
4.2. Standard Deviation of Measured Length Versus True Length	38
4.3. U.S. 20%-DAC Amplitude Drop Depth Measurements of Rough Flaws in Base Metal	49
4.4. Non-U.S. 20%-DAC Amplitude Drop Depth Measurements of Rough Flaws in Base Metal	49
4.5. U.S. 20%-DAC Amplitude Drop Length Measurements of Rough Flaws in Base Metal	50
4.6. Non-U.S. 20%-DAC Amplitude Drop Length Measurements of Rough Flaws in Base Metal	50
4.7. U.S. 20%-DAC Amplitude Drop Depth Measurements of Smooth Flaws in Base Metal	50
4.8. Non-U.S. 20%-DAC Amplitude Drop Depth Measurements of Smooth Flaws in Base Metal	50
4.9. U.S. 20%-DAC Amplitude Drop Length Measurements of Smooth Flaws in Base Metal	51
4.10. Non-U.S. 20%-DAC Amplitude Drop Length Measurements of Smooth Flaws in Base Metal	51

Tables

2.1. Summary of PISC-II Computer Files	4
2.2. List of Inspection Variables	5
2.3. Defining Categories for the Variable sizing Procedures	5
2.4. Summary of PISC-II Blocks	8
2.5. Material Variables	9
2.6. Summary of Grading Units in PISC-II Assemblies	12

2.7. Comparison of Detection Statistics for PNL Analysis Versus PISC-II Analysis	12
3.1. Summary of POD Fits to Flaw Depth	17
3.2. Summary of POD Fits to Flaw Length	18
3.3. Summary of POD Fits to Cross-Sectional Area	19
3.4. Detection Inspections by Block	20
3.5. Detection Inspections by Procedure	20
3.6. Regression Fit Comparison of Manual-20% DAC Procedures for U.S. Versus Non-U.S. Teams	21
3.7. Clad Versus Base-Metal POD	22
3.8. Goodness-of-Fit Statistics for Clad Versus Base Metal	24
3.9. Tests for Significance of Advanced Team Variation	25
3.10. Advanced Team PODs for Base-Metal Flaws	26
3.11. Advanced Team PODs for Clad Flaws	27
3.12. Summary of Team-to-Team Variability Tests	28
4.1. Summary of Sizing Inspections by Procedure and Country	40
4.2. Comparison of U.S./Non-U.S. 20%-DAC Amplitude Drop Depth Sizing Procedures	41
4.3. Summary of U.S./Non-U.S. Depth Sizing Performance	42
4.4. Comparison of U.S./Non-U.S. 20%-DAC Amplitude Drop Length Sizing Procedures	43
4.5. Summary of U.S./Non-U.S. Length Sizing Performance	44
4.6. Summary of Clad Versus Base-Metal Depth Sizing Performance	45
4.7. Summary of Clad Versus Base-Metal Length Sizing Performance	46
4.8. Comparison of Depth Sizing Regressions	47
4.9. Comparison of Length Sizing Regressions	48
5.1. Summary of Flaws Used in ASME-Type Clad Test	54
5.2. Summary of Embedded Flaws Used in ASME-Type Shell Weld Detection Test	54
5.3. Summary of Teams Passing ASME-Type Clad Detection Test	55
5.4. Summary of Teams Passing ASME-Type Clad Sizing Test	56
5.5. Summary of Passing Teams for ASME-Type Vessel Weld Detection Test	58

Executive Summary

This report provides an analysis of ultrasonic inspection performance data that was gathered during an exercise called PISC II (Programme for the Inspection of Steel Components, Phase 2), an international round robin test on ultrasonic inspection capability. Inspection data was gathered on four heavy section steel components which included two plates and two nozzle configurations. Teams from several European countries as well as the U.S. and Japan participated in this multi-year study. A total of 45 teams and approximately 429 inspections are listed in the computer data base that resulted from the PISC-II study.

The inspection data used in this study is a computerized data base available from PISC. Several extensive analyses have already been performed on this data base, and PISC-II has also summarized the results using simple statistical summaries. In this report, more formal statistical procedures have been applied to the computerized data.

Major conclusions from the analysis are:

- There is no significant difference between U.S. and other PISC-II participants using similar procedures.
- Advanced sizing procedures (e.g., SAFT, TOFT, ALOK, holography) characterize the flaws more accurately than amplitude drop methods.
- Team-to-team variability is a large factor in both detection and sizing performance. This finding illustrates the need for serious study of the human element in ultrasonic inspection.
- The detection performance demonstration requirements proposed by ASME Section XI clearly screen inspection capability well. As an example, only those teams that used special inspection techniques for the clad region had a high chance of successfully passing an ASME-type test. However, sizing techniques used by PISC-II participants would not have been successful in passing new sizing performance demonstration requirements adopted by ASME Section XI.

Acknowledgements

We would like to thank the staff at the Joint Research Centre in Ispra, Italy for their assistance in providing the PISC-II data base to us for analysis. This is a large data base, and it required extensive work to understand the data contained in it. The JRC staff headed by Serge Crutzen facilitated this process and were able to teach us the intricacies of this data base for which we are deeply grateful.

We would like to thank the NRC program manager, Dr. Joseph Muscara, for his support and guidance in conducting this re-evaluation. As always, the key to getting any manuscript published is due to the knowledge and coordination of Kay Hass in putting this report into the required format and correcting the many spelling, grammar, and punctuation errors.

1.0 Introduction

The U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research has funded a multi-year program at the Pacific Northwest Laboratory, entitled "Evaluation and Improvement in Nondestructive Examination Reliability for Inservice Inspection of Light Water Reactors." The objective of this work is to quantify the reliability of present U.S. inservice inspection techniques for light water reactor primary system components and to establish means for obtaining improvements in the reliability of inservice inspections. The PISC-II data base is by far the largest data base dealing with ultrasonic inspection capability of thick-section steel. Thus, this data base was analyzed to aid in quantifying the flaw detection and sizing capabilities for reactor pressure vessels.

This report provides an analysis of ultrasonic inspection performance data that was gathered during an exercise called PISC II (Program for the Inspection of Steel Components, Phase 2), an international round robin test on ultrasonic inspection capability. Inspection data was gathered on four heavy section steel components which included two plates and two nozzle configurations. Teams from several European countries as well as the U.S. and Japan participated in this multi-year study. A total of 45 teams and approximately 429 inspections are listed in the computer data base that resulted from the PISC-II study.

The PISC-I study grew out of a preceding study conducted by PVRC. The work performed in PISC-I was completed in the early 1980's and was to "investigate the ability of the 1974 PVRC procedure to detect, locate and size flaws in thick reactor steels by ultrasonic methods," (Nichols and Crutzen 1988). The general objective of the following study, PISC-II, was to "collect more data in order to obtain better-defined results and broader conclusions." More specific objectives identified for the PISC-II project included:

- To evaluate the effectiveness of NDT techniques either in use or being developed for the inservice inspection of reactor pressure vessel components, with regard mainly to detecting and characterizing service-induced flaws and fabrication flaws.
- To identify techniques for acceptance tests, preservice inspection (PSI) and inservice inspection which could be generally accepted.
- To "validate" the reliability of NDT techniques and procedures.
- To provide recommendations for national and international codes and standards.

The inspection data obtained from this study has been computerized, and is available to interested parties. Several extensive analyses have already been performed on this computerized data set, among them Murgatroyd, Crutzen, Vinch, and McFarlane 1988; Hemsworth 1986; and Davies 1987. The PISC-II Committee has also summarized the results using simple statistical summaries in (Nichols and Crutzen 1988). In this report, more formal statistical procedures have been applied to the computerized data set than those described in the PISC reports. A particular reason for the use of more formal statistical procedures (i.e., hypothesis tests, loglinear models, and regression models) is to more properly account for the experimental error in the measurements.

A major objective of this report is to extract information from the PISC-II data set that is relevant to inservice inspection in the U.S. Inspection capability is evaluated with the same sorts of statistical procedures employed in the piping inspection round robin and mini-round robin studies conducted at PNL (Heasler, et al. 1990; Doctor, Deffenbaugh, et al. 1988; Bates, Doctor, Heasler, and Burck 1987). Specific topics that are relevant to U.S. inspection capability include:

- Analysis of U.S. team performance with a comparison to overall performance.
- Analysis of team performance with respect to requirements of ASME Section XI, Appendix VIII for performance demonstration of detection capability and sizing accuracy.

This report contains a systematic analysis of the PISC-II data base. Section 2 provides a description of the PISC-II data base, the inspection variables and the materials variables that are involved, the scoring scheme that PNL will use in the data analysis, and

Introduction

comparison of the PNL-scored results with the PISC-II reported results. Section 3 concerns an analysis of detection performance and assesses the importance of the various variables considered in the analysis. Section 4 assesses the sizing performance of teams and procedures and defines the metric used to quantify the sizing performance. Section 5 contains an assessment of using the ASME Code Section XI Appendix VIII requirements to evaluate the procedures used in the PISC-II study. The final section contains a set of conclusions and recommendations that can be extracted from this analysis of the PISC-II data base.

2.0 Contents of the PISC-II Data Set

Large volumes of data are generated by an ultrasonic inspection. A typical inspection consists of several examinations of the material under investigation, each using a different transducer and viewing angle. The most basic data from such scans are recorded from oscilloscope responses and transducer locations, but data in this detail is not of interest in a study such as PISC-II. The inspectors are required to "reduce" the instrument responses into a summary that is a relevant description of material integrity. In other words, an NDE inspection is required to produce a list of potential flaws within the material.

For PISC-II, all ultrasonic inspectors were required to reduce their work to a list of "flaw" cuboids. The locations and dimensions of the cuboids were represented by a vector of six coordinates, which defined a three-dimensional rectangular volume in the material. This volume is called an "indication," to distinguish the inspection results from the actual flaws in the material.

The essential format of the computerized PISC-II data therefore consists of lists of indications, with each list representing the results of a single technique (e.g., 45° pulse echo). Important ancillary information is also present at the head of each list such as the block inspected, the technique used. In addition to information from individual inspection techniques, the PISC-II data (CLEAN.PROC) also contains the summary data that inspection teams provided and summary data developed by PISC-II computers that select the best results of each inspection technique. The summary data provided by each inspection team contains that team's best estimate of the condition of the test piece inspected (e.g., the location and size of all flaws defects in the test piece). The summary data provided by the PISC-II computers provided an optimized or upper bound estimate of NDE capability by selecting the best results from each individual technique.

2.1 Variables Recorded in the PISC-II Inspection Data Sets

The PISC-II data files contain "edited" and "raw" versions of the same inspection data. The "editing," which is called "cleaning" in PISC-II parlance, is basically designed to transform all indication

coordinates to one common coordinate systems. Teams tended to use their own special coordinate systems when recording the data. However other important differences also exist between the "cleaned" and "raw" data, the most important of them being:

- Coordinate calculation errors and transcription errors have been corrected.
- Many false call indications (i.e., indications not associated with an actual crack) have been eliminated from the cleaned data. The elimination of false call data from the PISC-II data base makes determination of false call probability impossible and seriously limits the utility of the data base.
- "Derivative" inspection results simulating 20%-DAC and 50%-DAC inspection have been produced from 10%-DAC results by a computer program.
- Derivative inspection results were produced by combining technique results with the aid of a computer program. These derivative results are referred to as the "best team" results in the PISC-II reports (Nichols and Crutzen 1988).

The complete PISC-II data set can be summarized as illustrated in Table 2.1. The inspection data consists of four files containing a total of 1508 redundant inspection records. In this report, we make use of only one of the data files, the "cleaned procedure" file, "CLEAN.PROC", which is the data set used in Nichols and Crutzen (1988) for most of their analysis. This file also contains the computer-compiled 20%-DAC and 50%-DAC inspections.

The data in the file CLEAN.PROC were compared with the results reported in Nichols and Crutzen (1988) to see if the computerized data was consistent with published results. Our test consisted of attempting to duplicate Table 1 of the PISC-II Report No. 5. We were able to duplicate the detection probabilities listed in Nichols and Crutzen with a few exceptions. After contacting the staff at the Joint Research Centre at Ispra, Italy, these discrepancies were resolved.

Table 2.1. Summary of PISC-II Computer Files

File	Number of Inspections in File	Contents
RAW.TECH	208	Original technique inspection results (DS6.1)
RAW.PROC	447	Original procedure inspection results (DS6.2)
CLEAN.TECH	391	"Cleaned" technique inspection results (DS6.1)
CLEAN.PROC	462	"Cleaned" procedure inspection results (DS6.2)
Total	1508	

2.1.1 Inspection Variables

Since the PISC-II data base contains data from each individual inspection, team summary data, and PISC-II computer summary data, analyzing the entire CLEAN.PROC data file would bias results because of the redundant versions of the inspection data. The analysis for this report uses only the team summary data. The team summary data was used because it was thought that data would more closely represent inspection performance related to inservice inspection. In order to analyze PISC-II data, PNL developed categories which describe procedures used for detection and sizing inspections performed on the PISC-II specimens. A summary of the detection and sizing categories used by PNL is provided in the following section.

The PISC-II round robin data files contain multi-page descriptions of each inspection procedure, and these were distilled into a set of consistently defined "inspection" variable values by the evaluators at the Joint Research Center. These inspection variables were included within the computerized data set and are the principal means for evaluating the role that different procedure variables have on inspection performance. A description of the detection and sizing categories that were used in this study is provided below in Table 2.2. The variables listed in this set are based upon the same information used in other PISC-II studies. However,

these variables, particularly the procedure variables, are constructed to allow us to answer the particular questions of interest and are unique to this study.

The variables in "CLEAN.PROC" were tersely coded, and it took some effort to assemble a complete description of the codes.

The variables listed above in the table are largely self-explanatory. "Team ID" contains a unique two-letter "blind" code that distinguishes between each team. "Country" separates the teams into U.S. and non-U.S. sets, so that it will be possible to evaluate the overall performance of U.S. teams, something that was forbidden for the PISC-II analysis. "Measurement Type" identifies whether or not inspections included the detection and sizing tasks. "Access" describes the surfaces on which the examinations were conducted. "Detection Procedure" categorizes the detection inspections. The notations 10%, 20%, 50% refer to the recording level specified in the procedure; "man" or "auto" identify the procedure as either manual or automatic; "adv" indicated that the procedure used imaging or signal processing. Examples of "adv" procedures are SAFT, holography, TOFT, or ALOK. Sizing Procedure categorizes the procedures used for sizing flaws. Table 2.3 provides a listing of the specific sizing techniques used in the "Sizing Procedure" category listed above.

Table 2.2. List of Inspection Variables

Variable	Possible Values
Team ID	AN, BA, BC, ...
Country	U.S., Non-U.S.
Measurement Type	Detection + Sizing, Detection Only, Sizing Only
Access	Unclad, Clad, Both
Detection Procedure	Advanced, Automated Procedure-10%, Automated Procedure-20%, Automated Procedure-50%, Manual Procedure-10%, Manual Procedure-20%, Manual Procedure-50%
Sizing Procedure	20% Amplitude Drop, 50% Amplitude Drop, Amplitude Drop Imaging, Diffraction, Focused Technique, Simple Imaging, Combination

Table 2.3. Defining Categories for the Variable sizing Procedures

Sizing Category	Description
20% Amplitude Drop, 50% Amplitude Drop, Amplitude Drop	<p>The sizing techniques in the amplitude drop category sized flaws by recording probe spatial information as outlined in early editions of the ASME Code. The techniques in this category include:</p> <ul style="list-style-type: none"> • 20% DAC Amplitude Drop Sizing Procedures • 50% DAC Amplitude Drop Sizing Procedures • Half Maximum Amplitude Procedure • Tandem at -6, -4, -2 dB • Amplitude Drop Procedure with unspecified drop • Phased Array
Computer Imaging	<p>This category includes any imaging system that used some type of signal processing in developing the image. These systems included:</p> <ul style="list-style-type: none"> • Digital Acoustic Holography • Line SAFT • 3D SAFT • Multifrequency Holography • Holography and SAFT Combination • 45° Probe Holography

Contents of PISC-II Data Set

Sizing Category	Description
Simple Imaging	<p>This category included imaging systems that simply displayed projection scans or "B", "C", "D" scans. The techniques were:</p> <ul style="list-style-type: none"> ● P-scan ● Interpretation of "B", "C", "D" scans
Combination	<p>This category literally used a combination of technologies to size flaws.</p> <ul style="list-style-type: none"> ● Combination of Focused Probes; SAFT; Tandem; TOF; Manual Scan ● Line Holography; Phased Array ● Focused SE Probes
Diffraction	<p>The sizing techniques in this category estimated flaw size by detecting ultrasonic signals from flaw tips. The techniques in this category included:</p> <ul style="list-style-type: none"> ● Time-of-Flight Spectroscopy ● Wide Angle Pulse Echo and Tip Diffraction ● Multiple Beam Satellite Pulse ● Time of Flight ● Spectroscopy ● TOFD and Pitch Catch ● ALOK and Time Integrated Method ● Tip Diffraction
Physically Focused Technique	<p>The sizing techniques in this category used exclusively focused beam technology to size flaws. The techniques were:</p> <ul style="list-style-type: none"> ● Dynamically Focused Probes ● -6 dB Focused Probe

During the PISC-II round robin tests, multi-page descriptions of each inspection were produced, and these were then distilled into a set of consistently defined "inspection" variables by the evaluators at JRC Ispra. These inspection variables were included in the computerized data set, and they provide the principal means for evaluating the role that equipment and procedures may have had on inspection performance. Among these variables, one identifier was available to distinguish between the data obtained by U.S. and non-U.S. teams. This has proven to be a useful feature during this reanalysis of the PISC-II data.

2.2 Material Variables: Description of PISC-II Blocks

The PISC-II round robin tests were conducted on four heavy section steel test specimens. This section provides a brief description of each test specimen and the variables that describe them. This section also provides a comparison between the variables used by PISC-II and the variables this report uses to describe the inspected material.

Test Plate No. 1 was provided by the Federal Republic of Germany. This was a flat plate with a longitudinal submerged arc weld that contained 15 intentional defects. This plate was 1045 mm x 1026 mm x 246 mm and was clad on one side in accordance with a German

specification. The defect types included: a) hot and cold microcracks, b) 20 mm deep macrocracks, c) slag inclusions, and d) porosity.

Plate No. 2 was provided by the United Kingdom. This plate was also flat and contained a submerged arc weld with a double "J" weld prep. This plate was 1540 mm x 1520 mm x 262 mm and was strip-clad on one side, and it contained the following fabrication flaws and service-induced defects: a) near-surface vertical cracks (generally 30-60 mm in height), b) welding flaws, c) slag inclusions, d) copper cracking, and e) carbon cracking.

Plate No. 3 was provided by Italy (JRC Ispra) and was a PWR inlet nozzle mock-up with weld flaws and service-induced defects. This plate was 2620 mm x 2300 mm x 250 mm. The nozzle O.D. was 1370 mm, nozzle I.D. was 780 mm, and the weld centerline diameter was 1530 mm. The weld was made using the submerged arc process, the plate was strip clad, and the nozzle was wire clad. Both simulated service-induced defects and welding defects were present. The service-induced defects included:

- a. 8 defects with diameters ranging from 3 to 25 mm
- b. 9 rectangular defects with heights ranging from 10 to 60 mm
- c. 4 sets of composite defects
- d. 3 machined notches at the nozzle inner radius

The welding defects included lack-of-fusion (4 defects), slag (5 defects), and 1 zone that had lack-of-clad bonding.

Plate No. 9 was provided by Japan. This plate also contained a nozzle with service-induced defects. The plate was 1950 mm x 1950 mm x 200 mm with a submerged arc weld. The nozzle O.D. was 535 mm, nozzle I.D. was 292 mm, and the weld centerline diameter was 890 mm. The plate was strip-clad and the nozzle was wire-clad. The defects in this plate included: a) 9 subsurface defects (10-15 mm deep); b) 3 near-surface defects (clad side, 10-15 mm deep, and surrounded by satellite defects); c) 5 near-surface defects (unclad side, 10-15 mm deep); and d) 3 fatigue cracks at the nozzle inner-radius.

The flaw descriptions present in the PISC-II data sets contained a simplified version of the information presented above; and in order to allow us to concentrate on the essential features in the round robin, we simplified the flaw descriptions even further. Table 2.4 lists the flaw categorizations that were actually employed on the data set used in this analysis.

The state of the four plates was summarized for this study by a set of flaw variables. Flaw variables are used to describe the condition of the material being inspected. This information was not available in the computer data set, but could be extracted from the PISC-II reports, particularly (Nichols and Crutzen, 1988). As one can see from the previous block descriptions, it is difficult to identify a single set of variables that are relevant for all four plates. In our attempt to construct just such a set of variables, our set is necessarily less detailed than PISC-II's description.

For the current work, the state of the inspected material was summarized with four discrete variables and flaw dimension which was divided into depth and length. Table 2.5 presents each variable and the possible values that could be associated with each variable. The classifications within the discrete variables have been simplified somewhat from their PISC-II counterparts. For example, in the PISC-II data set, the variable "location" identified up to seven locations within the block (i.e., within 18 mm of the inner surface, between 18 mm and 36 mm of the inner surface, etc.). We have simplified this variable to just three values: clad surface, mid-wall, outer-surface.

The material variables describe the state of units of material that we refer to as "grading units." These grading units are employed in the analysis to define detection. The grading units have been constructed so that each contains one flawed region or is blank; the flawed grading units are used to calculate Probability of Detection (POD) for a certain type and size of flaw, the "blanks" are used to calculate False Call Probability (FCP). The "grading units" employed in this study are plotted in Appendix A, along with the flaws that each contains.

Table 2.4. Summary of PISC-II Blocks

Block	Dimension, cm	Flaw Type			
		Multiple	Rough	Smooth	Volumetric
PLATE 1	#-flaws	1	3	1	3
	dx-minimum	0.8	0.4	1.1	0.3
	dx-maximum	0.8	2.8	1.1	1.0
	dy-minimum	99.7	6.0	8.4	9.6
	dy-maximum	99.7	9.3	8.4	47.5
	dz-minimum	1.5	0.5	1.5	0.3
	dz-maximum	1.5	1.4	1.5	0.7
PLATE 2	#-flaws	3	9	3	4
	dx-minimum	0.4	0.3	1.1	0.1
	dx-maximum	0.7	2.6	5.0	0.5
	dy-minimum	3.97	4.5	0.3	0.7
	dy-maximum	4.7	10.4	4.5	7.0
	dz-minimum	0.2	1.0	4.5	0.1
	dz-maximum	4.3	6.5	4.9	1.1
PLATE 3	#-flaws	NA	6	19	15
	dx-minimum	NA	0.1	0.1	0.2
	dx-maximum	NA	0.7	1.2	2.5
	dy-minimum	NA	1.0	0.3	0.4
	dy-maximum	NA	8.54	8.5	23.0
	dz-minimum	NA	0.3	0.3	0.1
	dz-maximum	NA	2.6	8.0	0.6
PLATE 3	#-flaws	1	8	5	3
	dx-minimum	0.1	0.5	0.3	0.1
	dx-maximum	0.1	3.9	0.8	0.2

Block	Dimension, cm	Flaw Type			
		Multiple	Rough	Smooth	Volumetric
	dy-minimum	0.7	4.4	1.0	0.6
	dy-maximum	0.7	8.5	6.6	1.1
	dz-minimum	0.6	0.8	0.4	0.3
	dz-maximum	0.6	3.4	2.3	0.6

Table 2.5. Material Variables

Variable	Possible Values
Block Identification	Plate 1, Plate 2, Nozzle 3, Nozzle 9
ASME Classification	Blank, Acceptable, Rejectable
Location	Clad, Mid-wall, Outer
Flaw Type	Blank, Multiple, Rough, Smooth, Volumetric
Flaw Dimension	(X1,Y1,Z1) to (X2,Y2,Z2)

2.3 Scoring: Calculation of Inspection Performance Statistics

There is general agreement on the two basic tasks that UT procedures must perform during inspection: flaw detection and sizing. Consequently, it is not surprising that inspection performance is quantified by statistics that fall into two broad categories; those that measure detection capability and those that measure sizing capability.

Detection capability is typically quantified by a probability or frequency, which represents the chances that an inspector will detect a flaw in a specified unit of (usually flawed) material. This statistic has many different names; PISC-II uses the terms "defect detection probability" and "defect detection frequency,"

while this report employs the terms "probability of detection" and "false call probability." Although the concept of probability of detection is relatively simple and extensively used, the method of its estimation is not, at least for the type of raw data produced by typical ultrasonic round-robin experiments.

This problem has such an effect on the results obtained that we have given it a special name -- the method of calculating detection (and sizing) statistics from indication cuboids is called "scoring." "Scoring" essentially involves the association of indication cuboids with flaws to determine whether or not the flaw was "detected." Difficulties in this simple objective occur because:

1. Location and sizing errors mean that the "real" indication produced by a flaw will have

- neither the same location nor shape as the true flaw.
2. There is not necessarily a one-to-one relationship between flaws and indications; one flaw may produce several indications and, conversely, several flaws may be represented by one indication.
 3. And finally, false calls can occur. Simply because an indication is close to a flaw, that does not mean it was produced by a signal from the flaw; it might be a mistake or it might originate from an innocuous condition.

Figures 2.1 through 2.3 illustrate these problems in two-dimensional form. In Figure 2.1, several indications are near an actual flaw. The nearest indication to the flaw is quite small. Should it be associated with the flaw? This would result in substantial sizing error. On the other hand, an indication that is closer to the size of the flaw exists a bit farther away. Perhaps this should be associated with the flaw? What criteria would be more reasonable for making the association, size or location?

In Figure 2.2, several indications have been randomly placed within the test object, indicating that the false call rate is very high. The black square represents a

Figure 2.1. Scoring with Multiple Indications

Figure 2.2. Scoring with High False Calls

very small flaw that was not detected. However, because the false call rate is so high, no point in the material is very far from an indication, including the small flaw. We may therefore conclude that small flaw was detected when it was simply near a false call.

A similar phenomenon can occur when the flaw density is high, as is the case in the PISC-II blocks. If flaws are oversized by the inspection procedure, an adjacent flaw may be responsible for detection of the flaw. For example, in Figure 2.3, one indication intersects two flaws, a large one and a small one. It seems reasonable to conclude that the large flaw was detected, but what about the small one? Did the inspector intentionally combine two flaws into one or did diffraction effects and tip signals cause the two flaws to look like one large flaw? If we conclude that the small flaw was detected, how can reasonable dimensions be assigned?

Scoring should be completely objective. It should be the result of the application of explicit rules to the data, with no ad-hoc exceptions. Our method of defining flaw detection consists of surrounding each flaw with a fairly large cuboid, called a grading unit. Detection is defined to occur whenever any indication intersects with the cuboid. Since the size of the grading unit chosen

Figure 2.3. Scoring with Multiple Flaws

can affect POD, the following guidelines are followed when laying out the grading units:

1. Make grading units large enough so that any reasonable location error would produce a detection.
2. Grading units should all be of commensurate size so that detection statistics can be compared.
3. Grading units should not be placed too closely to each other or other flaws; they should be far enough apart so that any reasonable location or sizing error could not affect an adjacent grading unit.
4. The defect(s) should be placed in the middle of the grading unit.
5. "Blank" grading units should be defined, so that the effect of false calls can be determined.

The grading units defined for this study are graphically displayed in Appendix A. During the construction of the grading units, we found it impossible to obey all the guidelines presented above. In particular, it was not

possible to use the same grading unit size to surround all flaws because of the large size of some flaws. Grading units also had to be squeezed too closely together in the blocks because of the high flaw density. Finally all "blank" grading units may not be perfectly flaw free (no piece of metal ever is). In this study, blank grading units were selected based on PISC-II data to ensure that they did not have any acoustical anomalies that exceeded 3 mm in size.

The grading unit characteristics are summarized in Table 2.6. Notice that the grading unit dimensions are quite consistent, except possibly for Plate 1. In Plate 1, one large flaw ran almost the entire length of the block, which required the use of one very long grading unit. This has resulted in an average "dy" dimension of 20.5 cm for this block.

2.4 Comparison of PNL-Scored Results to PISC-II Results

The following simple procedure was used to score data from the PISC-II round robin.

1. Detection is defined to occur whenever any indication intersects a grading unit.
2. When multiple indications intersect a grading unit, the largest indication is used to estimate flaw size.

Since our procedure for scoring the inspection results differs from that used by PISC, the question naturally arises: How different are our results from PISC's? Table 2.7 compares detection statistics for the two methods of scoring. The table lists the number of inspections/flaws that were classified as detected or missed by the two scoring methodologies.

Notice that about 90% of the classifications agree between the two methods of scoring. Generally, our method of scoring is slightly more conservative than PISC's; PISC-II classifies 372 flaws as detected that PNL classifies as missed, while PNL classifies only 70 as detected that they classified as missed. This is surprising because we surround the flaw with a fairly large grading unit, while they use a tolerance box that is generally smaller than our grading unit. Hence, we

Table 2.6. Summary of Grading Units in PISC-II Assemblies

	Plate #1	Plate #2	Nozzle #3	Nozzle #9
Number of Blank Grading Units	7.0	13.0	11.0	12.0
Number of Flawed Grading Units	8.0	19.0	40.0	17.0
Average Dimensions (cm)				
dx	10.0	5.2	5.0	5.0
dy	20.5	8.0	8.3	7.8
dz	4.4	5.3	5.0	4.9

Table 2.7. Comparison of Detection Statistics for PNL Analysis Versus PISC-II Analysis

PNL	PISC-II		Total
	Missed	Detected	
Missed	2112	372	2484
Detected	70	2574	2644
Total	2182	2946	5128

would have expected that the larger scoring box would give the lowest miss rate. This illustrates the fact that in the PISC-II analysis, a subjective scoring scheme was employed in which a PISC-II team reviewed the data and selected a team's indication to be scored with each flaw.

The scoring algorithm will also affect the indicated sizes that are associated with the flaws. Approximately 95% of flaw depths and lengths agree between the PISC-II and PNL method of scoring. Of the 5% that do disagree, the differences can be substantial, particularly in Plate 1. Figures 2.4 and 2.5 display the differences between our scored sizes and PISC's. (These plots only contain the 5% of the data that disagrees.)

Since we assign the size of the largest intersecting indication to a flaw, one would expect our flaw sizes to be larger than PISC's and the plots do display such a

feature. In the case of flaw lengths, the results are greatly affected by the flawed area in Plate 1 that is 100 cm long. The other flaws in the block are close enough to this large flaw to have indications from the large flaw intersect with the smaller flaws. For a case such as this, scoring becomes a fundamentally ambiguous procedure. Therefore, Plate 1 was not used in this analysis.

**Figure 2.4. Plot of Flaw Lengths for PNL
Versus PISC-II Scored Data**

**Figure 2.5. Plot of Flaw Depths for PNL
Versus PISC-II Scored Data**

3.0 Analysis of Detection Performance

This section compares the inspection capability of U.S. teams with non-U.S. teams and analyzes the effect of flaw location (within clad versus base-metal), variability of teams, and the effectiveness of particular classes of procedures. To perform formal tests to determine the effects of the variables and modeling procedures, Chi-squared and F-test statistics are employed. The data employed for these calculations consists of essentially all summary inspections in the "cleaned procedure" file, excluding those that were solely done for sizing, and those originating from Plate 1. (This block was eliminated because of the ambiguity in the data caused by the long flaw that made scoring the data very difficult.)

When inspection procedures are evaluated under many different sets of conditions, one of the most important problems that must be dealt with is the problem of confounding. Confounding occurs when the effect of one variable can be mistaken for the effect of another. Unless the experiment is carefully designed, one can not easily distinguish the effect of one variable from another; in the worst case, all variables must be considered simultaneously to obtain a true picture of their effects. For example, a curve that plots POD (i.e., the number of detections/total number of inspections of a given flaw) versus crack depth may not express the true effect of depth on POD if the large cracks had a large crack opening and the small cracks were tight. In this case, crack size is said to be confounded with crack tightness and, it would be impossible to determine whether the observed pattern in the POD's was due to crack size or crack tightness.

In this analysis, we have attempted to "control for" the most important variables in the data set to account for confounding problems. Unfortunately, there are too many different conditions present in the PISC data to allow us to take into account all potential confounding problems. When producing POD curves, we always consider the following variables: procedure type, flaw type, flaw location, and of course, flaw size.

All POD curve analyses in this section will utilize flaw depth (measured in cm) as the flaw size metric, because flaw depth is the one flaw dimension that is most closely related to the flaw's importance or propensity to cause failure. It is also possible to conduct POD regressions using other size metrics such

as length, area, volume, etc. In Appendix B, POD curves involving flaw length are also presented for the readers' reference.

3.1 POD Regression Methodology

POD curves are fit to binary detection data in this report using quantal response regression (McCullagh and Nelder 1983) a form of weighted regression. This regression can be used to fit a parametric model to POD, which relates POD to other variables included in the model. In this way, the effect of the variable(s) on POD can be gauged.

For this work, we have chosen a "logistic" model for the regression function, which produces a general, S-shaped curve as illustrated in Figure 3.1. Logistic regression attempts to fit a function with the form:

$$POD(s_i) = \text{logit}(\beta_1 + \beta_2 s_i)$$

where s_i represents the size of flaw i and the logit function is defined by:

$$\text{logit}(z) = [1 + \exp(-z)]^{-1}$$

Since flaw size should have a very strong relationship to POD, logistic regression should explain a major part of the variability in POD. A typical result from a logistic regression is illustrated in Figure 3.1. The logistic curve is fit to data obtained from U.S. teams who have used a Manual-20% procedure and inspected rough cracks in base material. This particular logistic regression uses depth (cm) as the crack size metric. It is important to note that each point in the plot represents the observed POD obtained on a particular flaw (i.e., the number of detections/total number of inspections on the flaw). Thus the scatter of the points around the POD curve represents binomial variability AND flaw to flaw variability.

Certain features of this regression are worth explaining. First of all, note that there are many flaws with "0" length in the data set; these are blanks. Since some teams have indicated that "flaws" exist in the blanks, the

Analysis of Detection Performance

POD curve does not intercept the origin and $POD(0)$ is positive. In fact, $POD(0)$ is equivalent to false call probability.

Secondly, the data from the POD-curve fits almost always contain more variation than that which can be attributed to binomial variation, and this tends to distort regression test statistics and confidence bounds. The POD plots indicate the extra variation is chiefly the result of flaw-to-flaw variability, and we have attempted to account for this feature in the data by including flaw-to-flaw variability in confidence bound and test statistic calculations. There are several methods to account for this variability, including some fairly complex random effect models. This analysis utilized a method described in McCullagh and Nelder (1983) which is quite simple. The more complex random effects models would require more extensive data than that available in the PISC data base, and thus we have employed a more simple method in this report.

In Figure 3.1, data points that show anomalous variation from the regression line are indicated with an "H" (for high) or an "L" (for low). In the example, there are several small flaws that exhibited an anomalously high POD, and two large flaws that exhibited an anomalously low POD.

Another extremely useful product of a regression fit are the associated confidence bounds and test statistics that allow us to determine the accuracy of the fit and intelligently compare it to other fits. In the example presented in Figure 3.1, the POD curve has been surrounded by 95% confidence bounds.

Appendix B presents a complete set of POD curve fits for the most relevant procedures (to ASME and U.S. performance; i.e., advanced and the manual procedures). Multiple flaws have been omitted from these regressions because too little data exists for these flaws. These fits relate POD to flaw length and depth.

We have used POD regression to summarize the effect of using a flaw size metric other than depth in the analysis. Tables 3.1 through 3.3 list comparable POD regression results using depth, length, and cross-sectional area of the flaws.

In these tables, the estimated POD is given for three flaw sizes; a flaw of size 0, (which is the false call probability); a median-sized flaw; and a flaw at the upper quantile in size. More relevant to our concerns are the goodness-of-fit (GOF) statistics which can be used to compare the three flaw size metrics to each other. These goodness-of-fit statistics determine whether the logistic regression satisfactorily explains the structure within the POD data.

If the variation were purely binomial, the goodness-of-fit statistic should be approximately equal in magnitude to 1 (actually χ^2/DOF where DOF = degrees of freedom) while the expected level of significance (LOS) should be 0.50. Because of the flaw-to-flaw variability in the data, all the goodness-of-fit statistics are too large. Nevertheless, these statistics can be usefully used to determine which size metric is most closely related to POD.

Ordering the three size variables by GOF, one finds that depth is most effective in explaining POD (with $\chi^2/DOF=6.4$), that length is next and only slightly worse with $\chi^2/DOF=6.5$, and finally cross-sectional area is last with a value of 7.1. Linear combinations of both depth and length were also examined, but this regression did not result in an equation that fit the POD data significantly better or was more reasonable than the three size variables examined above. From these goodness-of-fit statistics, one can conclude that either depth or length would be a reasonable size metric to use for this data.

3.2 Comparison of U.S. with Non-U.S. Teams

This section compares the detection performance of U.S. teams with the other teams that participated in PISC. There are two reasons for making such a comparison. First of all, it is natural to ask how well the U.S. teams and procedures compare internationally -- are they better or worse than other teams and, if so, why? Secondly, we would like to use the whole PISC data set to evaluate questions concerning U.S. inspection capabilities, if the data is indeed applicable. Because the data set is much more extensive than the U.S. data alone, use of the complete data set will

Table 3.1. Summary of POD Fits to Flaw Depth

Flaw		POD (Depth, cm)					
Location	Type	POD(0)	POD(1.2)	POD(4.6)	GOF	DOF	LOS
Advanced		Procedures					
Base	Rough	0.1634	0.5711	1.00	6.0	52	0.0E+00
Base	Smooth	0.1437	0.2331	0.62	3.2	42	0.0E+00
Base	Volumetric	0.1228	0.9927	1.00	3.3	46	0.0E+00
Clad	Rough	0.0471	0.4233	1.00	1.7	13	6.1E-02
Clad	Smooth	0.1302	0.2603	0.80	2.0	25	2.4E-03
Clad	Volumetric	0.0800	0.9733	1.00	2.2	18	2.0E-03
Manual-10% DAC		Procedures					
Base	Rough	0.0914	0.8349	1.00	8.2	52	0.0E+00
Base	Smooth	0.1046	0.1995	0.68	3.4	42	0.0E+00
Base	Volumetric	0.0482	0.9998	1.00	2.4	46	4.8E-07
Clad	Rough	0.1022	0.3716	0.98	5.8	13	0.0E+00
Clad	Smooth	0.1521	0.2922	0.81	2.9	25	2.4E-06
Clad	Volumetric	0.0994	0.9753	1.00	3.6	18	4.2E-07
Manual-20% DAC		Procedures					
Base	Rough	0.0825	0.6854	1.00	27.9	52	0.0E+00
Base	Smooth	0.0931	0.1399	0.37	7.6	42	0.0E+00
Base	Volumetric	0.0254	0.9993	1.00	3.4	46	0.0E+00
Clad	Rough	0.0463	0.3952	1.00	2.3	13	4.0E-03
Clad	Smooth	0.0925	0.1710	0.60	4.2	25	0.0E+00
Clad	Volumetric	0.0860	0.4367	1.00	5.1	18	0.0E+00
Total					6.4	588	

Table 3.2. Summary of POD Fits to Flaw Length

Flaw		POD (Length, cm)					
Location	Type	POD(0)	POD(4.6)	POD(8.53)	GOF	DOF	LOS
Advanced		Procedures					
Base	Rough	0.132	0.66	0.95	3.0	52	0.0E+00
Base	Smooth	0.121	0.52	0.86	3.0	42	0.0E+00
Base	Volumetric	0.136	0.57	0.89	8.6	46	0.0E+00
Clad	Rough	0.046	0.99	1.00	3.0	13	1.7E-04
Clad	Smooth	0.203	0.60	0.87	2.8	25	3.0E-06
Clad	Volumetric	0.145	0.30	0.49	32.4	18	0.0E+00
Manual-10% DAC		Procedures					
Base	Rough	0.063	0.83	0.99	4.0	52	0.0E+00
Base	Smooth	0.067	0.58	0.95	2.5	42	4.2E-07
Base	Volumetric	0.080	0.63	0.96	3.0	46	0.0E+00
Clad	Rough	0.085	0.91	1.00	3.2	13	7.6E-05
Clad	Smooth	0.230	0.54	0.80	3.8	25	0.0E+00
Clad	Volumetric	0.102	0.48	0.85	2.6	18	1.7E-04
Manual-20% DAC		Procedures					
Base	Rough	0.042	0.72	0.99	7.59	52	0.0E+00
Base	Smooth	0.061	0.34	0.76	4.8	42	0.0E+00
Base	Volumetric	0.050	0.33	0.77	15.9	46	0.0E+00
Clad	Rough	0.073	0.92	1.00	21.9	13	0.0E+00
Clad	Smooth	0.167	0.32	0.49	6.4	25	0.0E+00
Clad	Volumetric	0.098	0.13	0.16	3.7	18	1.2E-07
Total					6.5	588	

Table 3.3. Summary of POD Fits to Cross-Sectional Area

Flaw		POD (Area, cm ²)					
Location	Type	POD(0)	POD(5.94)	POD(27.6)	GOF	DOF	LOS
Advanced		Procedures					
Base	Rough	0.1887	0.409	0.97	7.5	52	0.0E+00
Base	Smooth	0.1523	0.205	0.49	3.3	42	0.0E+00
Base	Volumetric	0.1567	0.849	1.00	3.6	46	0.0E+00
Clad	Rough	0.0753	0.539	1.00	4.7	13	6.0E-08
Clad	Smooth	0.2130	0.383	0.93	2.6	25	1.5E-05
Clad	Volumetric	0.1478	0.652	1.00	3.74	18	1.2E-07
Manual-10% DAC		Procedures					
Base	Rough	0.1173	0.622	1.00	8.5	52	0.0E+00
Base	Smooth	0.1168	0.172	0.52	4.0	42	0.0E+00
Base	Volumetric	0.0917	0.994	1.00	2.9	46	0.0E+00
Clad	Rough	0.1356	0.407	0.99	4.3	13	2.4E-07
Clad	Smooth	0.2588	0.383	0.84	3.9	25	0.0E+00
Clad	Volumetric	0.1270	0.975	1.00	2.8	18	5.5E-05
Manual-20% DAC		Procedures					
Base	Rough	0.1025	0.496	1.00	25.1	52	0.0E+00
Base	Smooth	0.1065	0.129	0.24	8.8	42	0.0E+00
Base	Volumetric	0.0597	0.848	1.00	8.0	46	0.0E+00
Clad	Rough	0.0593	0.654	1.00	4.5	13	6.0E-08
Clad	Smooth	0.1732	0.239	0.58	6.2	25	0.0E+00
Clad	Volumetric	0.0964	0.249	0.95	4.3	18	0.0E+00
Total					7.1	588	

Table 3.4. Detection Inspections by Block

Block #	Country	
	Non-U.S.	U.S.
Plate 1	50	0
Plate 2	70	5
Nozzle 3	39	10
Nozzle 9	66	13
Total	225	28

Table 3.5. Detection Inspections by Procedure

Procedure	Country	
	Non-U.S.	U.S.
Advanced	73	3
Automated-10% DAC	20	0
Automated-20% DAC	11	0
Automated-50% DAC	14	0
Manual-10% DAC	32	0
Manual-20% DAC	50	21
Manual-50% DAC	25	4

provide a much more complete picture of inspection capability than provided by the U.S. teams alone. The U.S. teams did not inspect all the blocks and did not use the wide variety of procedures as employed by other teams that participated in the PISC program. U.S. teams generally concentrated on ASME procedures and did not inspect Plate 1. Tables 3.4 and 3.5 contrast U.S. participation to non-U.S. teams. As one can see from the tables, U.S. teams produced about

11% of the data in PISC, with the bulk of the inspections using an ASME, Manual-20% procedure. In order to produce a reasonable comparison of U.S. with non-U.S. teams, we have again eliminated Plate 1 inspections, because data from it is biased by one excessively large flaw and the U.S. teams did not inspect it. For the same reason, we are restricting comparisons to teams using Advanced and Manual-20% procedures. Because there is relatively little U.S. data in the Advanced procedure category, we can not expect a very accurate comparison for these procedures. The most stringent comparison of U.S. versus non-U.S. performance is accomplished using Manual-20% procedures.

Figures 3.1 to 3.12 present POD curve fits for the U.S. versus non-U.S. team who used Manual-20% procedures. In performing these fits, we have controlled for flaw type (volumetric, rough, smooth) and location (base, clad). Therefore, we have produced a separate POD curve for each set of flaw type/location conditions.

In most cases, the U.S. and non-U.S. POD curves correspond strongly to each other. The cases that display the greatest difference are the POD curves for smooth flaws in base metal (Figures 3.3 and 3.4). In this case, the U.S. POD curve is below the non-U.S. curve, but not greatly so, as indicated by the confidence bounds. For both sets of teams, the POD data shows the same patterns; there is a cluster of moderate-sized flaws around 2 cm in depth that are consistently detected, but two larger flaws (of size 5 cm and 6.5 cm) which are poorly detected. There is, therefore, a good deal of flaw-to-flaw variability in this data. This variability contributes to the size of the confidence bounds surrounding the POD-curve estimates.

It is possible to conduct a formal test to determine whether or not U.S. and non-U.S. performance is the same. To accomplish this, one can calculate an F-statistic (Appendix C) that measures how well a common POD curve would fit both U.S. and non-U.S. data. If the F-statistic is large (the expected value of the statistic is 1 when there is no difference between the two sets of teams), then U.S. and non-U.S. data are "significantly" different. Table 3.6 presents the

Table 3.6. Regression Fit Comparison of Manual-20% DAC Procedures for U.S. Versus Non-U.S. Teams

Condition	DOF ₁	DOF ₂	F-Stat	LOS
Base Metal, Rough Flaw	2	90	0.49	0.61
Base Metal, Smooth Flaw	2	74	0.59	0.56
Base Metal, Volumetric	2	82	7.74	0.00
Clad, Rough Flaw	2	24	0.14	0.87
Clad, Smooth Flaw	2	48	0.29	0.75
Clad, Volumetric	2	30	2.49	0.10
Total Manual-20% DAC	12	348	0.92	0.53

F-statistics for the different flaw types/locations inspected by the Manual-20% and advanced procedures.

The last column in the table presents the level of significance (or P-value) associated from each test statistic.

The last row in the table presents a summary statistic for all flaw types and locations and can be used to determine whether there is generally any difference between Manual-20% U.S. versus Manual-20% non-U.S. teams. From the level of significance recorded in this table (if LOS is small, then the variables are different; and if it is large, then there is no difference), one can conclude that U.S. performance is not significantly different from non-U.S. performance.

We will make use of this conclusion in the remainder of POD analysis presented in this section. In the other analyses presented in this section, we will combine U.S. data with non-U.S. data, and assume that non-U.S. data is representative of U.S. performance. This of course, is an extrapolation, because the PISC-II data can not tell us how U.S. versus non-U.S. automatic procedures compare or whether the PISC-II Manual-10% performance would be representative for the U.S.

Three U.S. teams also employed advanced procedures during the round robin, and these can also be compared to their non-U.S. counterparts, although the

comparisons are much less accurate. Figures 3.13 to 3.24 present the advanced procedure comparisons. From the POD curves presented in these data, there is no conclusive evidence that the U.S. teams are better or worse than the other advanced teams. There is insufficient data for a formal comparison.

3.3 Comparison of CLAD POD with Base-Metal POD

Another important question that can be answered with this data set is: What is the difference between detection of defects in the clad and the base metal?

Table 3.7 provides a simple comparison of clad POD with that in the base metal. The reader should be warned that the comparisons derived from this table are necessarily crude, because they do not take into account differences in flaw size between the two sets of flaws. The flaws within the clad region are, on average, a bit smaller than those in base metal. In the base-metal flaw population, a few of the flaws are much larger than any in the clad population. In the table, only the most sensitive procedures are listed (i.e. 10%-DAC and 20%-DAC), as they provide the best comparison.

Inspection of Table 3.7 shows that the POD values for the clad/base metal region are uniformly less than

Table 3.7. Clad Versus Base-Metal POD

Procedure		Flaw Type			
		Blank	Rough	Smooth	Volumetric
Advanced					
Base Metal	POD	0.04	0.89	0.61	0.60
	NOBS*	197	111	38	67
Clad	POD	0.00	0.94	0.55	0.32
	NOBS	50	52	94	50
Automated-10% DAC					
Base Metal	POD	0.06	0.94	0.79	0.74
	NOBS	115	65	34	53
Clad	POD	0.03	0.93	0.69	0.42
	NOBS	30	28	72	36
Automated-20% DAC					
Base Metal	POD	0.00	0.82	0.67	0.50
	NOBS	58	34	18	26
Clad	POD	0.00	0.67	0.22	0.12
	NOBS	14	12	36	17
Manual-10% DAC					
Base Metal	POD	0.03	0.94	0.66	0.58
	NOBS	258	154	62	90
Clad	POD	0.00	0.9	0.55	0.36
	NOBS	58	52	132	70
Manual-20% DAC					
Base Metal	POD	0.01	0.91	0.53	0.40
	NOBS	605	370	163	223
Clad	POD	0.01	0.91	0.36	0.18
	NOBS	129	101	318	169

*NOBS = number of observations

those associated with flaws in the base metal. Upon more detailed analysis, one finds that not all inspection procedures used to inspect clad flaws were specifically designed to detect flaws at the clad/base-metal interface. Some inspection procedures used minimum ASME Code requirements, which at that time (≈ 1983) only specified using 45° and 60° shear wave search units. Other teams used inspection procedures which were specifically designed to detect flaws at the clad/base-metal interface (e.g., 70° RTD search units). Procedures that were specifically designed to inspect the clad/base-metal region performed well. Section 5 provides a more detailed analysis to prove this point.

Table 3.7 also allows false calls to be compared. One would expect the false call probability within the clad to exceed that obtained within base metal, because cladding is a noisier material. However the values in the table show that this is not the case. False call probability (i.e., POD(blank)) seems to be about five times higher in base metal than in the clad region. (One reason for the increased false call probability in the base metal region may be sound beam partitioning and redirection in the clad which cause indications to appear as though originating in the base metal. Another explanation for the increased false call probability in the base metal may be that the improved performance of POD in base metal was achieved at the expense of increased false calls. A third is that it may be a result of how false calls were cleaned from the data by the PISC-II Data Analysis Group.)

Another way to compare PODs obtained from the two regions is with the use of POD curves. This type of analysis automatically takes the effect of flaw size into effect, which offers some advantage to the POD table presented above. In Figures 3.25 to 3.27, fitted POD-curves for the Manual-20% DAC procedures are compared. The curves are surrounded by 95% confidence bounds, and they are plotted over the range of data used in their fit. The flaw "size" used in these fits was depth, measured in cm. Relevant information concerning the range is obvious from the figures; the base-metal flaws in the data set include larger flaws than in clad and consequently provide more POD information on larger flaws.

From the figures, there is an indication that base-metal POD is better than clad POD, particularly for

volumetric flaws. This conclusion is also supported by regression test statistics designed to compare clad with base metal POD as presented in Table 3.8.

Except for the false call probabilities, which are surprisingly consistent, the tables display large team-to-team differences of POD within the columns. From the Chi-square statistics given in Table 3.9, one can determine that this variability is indeed significant. The Chi-square statistics measure the variability between teams.

3.4 Team-to-Team Variability

POD tables and Chi-square statistics can also be used to evaluate team-to-team variability. In Tables 3.10 and 3.11, POD statistics are calculated for all advanced teams who inspected base metal flaws and underclad flaws. Since the category of "Advanced Procedures" includes a grab bag of ultrasonic imaging techniques (SAFT, Holography, Siamese Imaging, etc.), one would expect sizeable differences between teams.

Table 3.12 summarizes the situation for all different classes of procedures in the data set. The formal Chi-squared test shows that same results observed for the advanced procedures; namely, that team-to-team variability is quite significant, indicating that the detection data contains a great deal of team-to-team variability.

Another interesting feature in the table are the relatively high Chi-squared values associated with the manual procedures. Even though these teams used the same amplitude recording criteria (20% or 50% DAC) to investigate indications, the large variation in POD indicates that the decision criteria for determining "defective" versus "nondefective" indications was not solely based upon amplitude. Additional "human factors" were also influencing the decisions.

3.5 What Procedures were the Best?

The POD curve results can also be used to determine the "best" procedure type. A "good" POD curve is one that exhibits a low false call probability (POD(0) is near

Table 3.8. Goodness-of-Fit Statistics for Clad Versus Base Metal

Condition	DOF ₁	DOF ₂	F-Stat	LOS
Advanced				
Rough	2	57	1.16	0.32
Smooth	2	61	0.36	0.70
Volumetric	2	56	1.29	0.28
Manual-10% DAC				
Rough	2	57	4.97	0.01
Smooth	2	61	0.40	0.68
Volumetric	2	56	5.41	0.01
Manual-20% DAC				
Rough	2	57	1.07	0.35
Smooth	2	61	0.80	0.45
Volumetric	2	56	11.21	0.00
Total	18	522	2.69	0.00

0) and a large POD for flaws that fracture mechanics analysis shows to be important. In fact, the "ideal" POD curve resembles a step function. When performing such rankings with POD curves, one must be aware of the potential problems that could be caused by team-to-team and flaw-to-flaw variability. The confidence bounds computed from the logistic regressions presented in the figures may be too optimistic.

Figure 3.28 presents the POD curves for the detection of rough flaws in base metal. It is interesting to note that the class with the highest POD is Manual-10% DAC and not the advanced procedures. For the manual procedures, the POD curves behave in an understandable manner, with Manual-10% DAC displaying the highest POD and Manual-50% DAC the lowest. The Manual-10% DAC, Manual-20% DAC, and Advanced procedures are not significantly different (due to measurement error), but this group is

significantly better than the Manual-50% DAC procedure.

These two clusters of procedures are not significantly different within themselves, but are significantly different between themselves.

In Figure 3.29, the same general pattern exists between the procedures for smooth flaws as for rough flaws. The manual procedures are correctly ordered according to sensitivity; however, in this set of conditions, the advanced procedures basically tie with the Manual-10% DAC as the best.

3.6 Summary

This section provides a brief summary of the more important observations concerning detection capability that resulted from this analysis.

Table 3.9. Tests for Significance of Advanced Team Variation

Flaw Type		Base Metal	Clad
Blank	Chi ²	226	56
	DOF	6	7
	LOS	0.000	0.000
Rough	Chi ²	92	40
	DOF	6	6
	LOS	0.000	0.000
Smooth	Chi ²	40	116
	DOF	5	6
	LOS	0.000	0.000
Volumetric	Chi ²	27	25
	DOF	6	7
	LOS	0.000	0.001

- Using a logistic model, it was determined that flaw depth was the most effective parameter in explaining POD, that flaw length was next, and that cross-sectional area was the least effective.
- Using a formal F-statistic test, it was determined that the detection capability of U.S. teams participating in PISC II was not significantly different than non-U.S. teams.
- Using a formal Chi-squared test, it was determined that the PISC-II data contains a great deal of team-to-team variations. Even teams that used the same amplitude recording level (20% or 50% DAC) showed great variability. This finding indicates that in addition to flaw response amplitude, "human factors" also greatly influences decisions on flaw detection.
- Using the logistic model, it was determined that Manual 20%-DAC procedures were significantly better than Manual 50%-DAC procedures. 10%-DAC and Advanced procedures performed as well as and, in some instances, better than 20%-DAC procedures.

Table 3.10. Advanced Team PODs for Base-Metal Flaws

Team		Flaw Type			
		Blank	Rough	Smooth	Volumetric
BA	POD	0.000	0.947	0.444	0.462
	NOBS*	35	19	9	13
IS	POD	0.000	0.923	0.000	0.667
	NOBS	21	13	2	3
LN	POD	0.239	0.950	0.857	0.909
	NOBS	92	40	7	22
SM	POD	0.063	0.857	NA	0.0
	NOBS	16	7	0	2
SR	POD	0.000	1.000	1.000	0.700
	NOBS	8	4	7	10
SW	POD	NA	NA	NA	NA
	NOBS	0	0	0	0
TH	POD	0.031	0.667	0.000	0.250
	NOBS	32	21	4	4
YC	POD	0.069	0.882	0.667	0.308
	NOBS	29	17	9	13

*NOBS = number of observations

Table 3.11. Advanced Team PODs for Clad Flaws

Team		Flaw Type			
		Blank	Rough	Smooth	Volumetric
BA	POD	0.000	0.857	0.158	0.100
	NOBS*	7	7	19	10
IS	POD	0.000	1.000	0.400	0.500
	NOBS	4	4	10	4
LN	POD	0.000	0.962	0.889	0.571
	NOBS	26	26	36	28
SM	POD	0.000	1.000	0.250	0.200
	NOBS	4	5	4	5
SR	POD	0.000	1.000	0.833	0.800
	NOBS	3	2	12	5
SW	POD	1.000	NA	NA	1.000
	NOBS	1	0	0	2
TH	POD	0.000	0.750	0.556	0.000
	NOBS	5	4	3	4
YC	POD	0.000	1.000	0.556	0.222
	NOBS	7	6	18	9

*NOBS = number of observations

Table 3.12. Summary of Team-to-Team Variability Tests

Procedure	Chi ²	DOF	Chi ² /DOF GOF Between Teams
Advanced	658	59	11.2
Automated-10% DAC	413	50	8.3
Automated-20% DAC	218	25	8.7
Automated-50% DAC	322	35	9.2
Manual-10% DAC	821	57	14.4
Manual-20% DAC	2005	128	15.7
Manual-50% DAC	890	19	46.8
Total	5328	373	14.3

Figure 3.1. U.S. POD, Base Material, Rough Crack, Manual-20% DAC Procedure

Figure 3.2. Non-U.S. POD, Base Material, Rough Crack, Manual-20% DAC Procedure

Figure 3.3. U.S. POD, Base Material, Smooth Crack, Manual-20% DAC Procedure

Figure 3.5. U.S. POD, Base Material, Volumetric Crack, Manual-20% DAC Procedure

Figure 3.4. Non-U.S. POD, Base Material, Smooth Crack, Manual-20% DAC Procedure

Figure 3.6. Non-U.S. POD, Base Material, Volumetric Crack, Manual-20% DAC Procedure

Analysis of Detection Performance

Figure 3.7. U.S. POD, Clad Material, Rough Crack, Manual-20% DAC Procedure

Figure 3.9. U.S. POD, Clad Material, Smooth Crack, Manual-20% DAC Procedure

Figure 3.8. Non-U.S. POD, Clad Material, Rough Crack, Manual-20% DAC Procedure

Figure 3.10. Non-U.S. POD, Clad Material, Smooth Crack, Manual-20% DAC Procedure

Figure 3.11. U.S. POD, Clad Material, Volumetric Crack, Manual-20% DAC Procedure

Figure 3.13. U.S. POD, Base Material, Rough Crack, Advanced Procedure

Figure 3.12. Non-U.S. POD, Clad Material, Volumetric Crack, Manual-20% DAC Procedure

Figure 3.14. Non-U.S. POD, Base Material, Rough Crack, Advanced Procedure

Analysis of Detection Performance

Figure 3.15. U.S. POD, Base Material, Smooth Crack, Advanced Procedure

Figure 3.17. U.S. POD, Base Material, Volumetric Crack, Advanced Procedure

Figure 3.16. Non-U.S. POD, Base Material, Smooth Crack, Advanced Procedure

Figure 3.18. Non-U.S. POD, Base Material, Volumetric Crack, Advanced Procedure

Figure 3.19. U.S. POD, Clad Material, Rough Crack, Advanced Procedure

Figure 3.21. U.S. POD, Clad Material, Smooth Crack, Advanced Procedure

Figure 3.20. Non-U.S. POD, Clad Material, Rough Crack, Advanced Procedure

Figure 3.22. Non-U.S. POD, Clad Material, Smooth Crack, Advanced Procedure

Analysis of Detection Performance

Figure 3.23. U.S. POD, Clad Material, Volumetric Crack, Advanced Procedure

Figure 3.25. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Rough Flaws

Figure 3.24. Non-U.S. POD, Clad Material, Volumetric Crack, Advanced Procedure

Figure 3.26. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Smooth Flaws

Figure 3.27. Clad Vs. Base Metal Material POD for Manual-20% DAC Procedure on Volumetric Flaws

Figure 3.29. Comparison of Procedures for Smooth Flaws in Base Metal

Figure 3.28. Comparison of Procedures for Rough Flaws in Base Metal

4.0 Sizing Performance

After flaws have been detected, it is necessary to accurately determine their size, particularly if the procedure is very sensitive and can detect small flaws that may have no significance to structural integrity. In the PISC-II study, flaw sizing consists of specifying a "cuboid" thought to contain the prospective flaw. From destructive evaluation, the "true" cuboids that enclose actual flaws are available for comparison. Sizing performance is described by a set of regression equations that relate the observed flaw size to the true size.

In this section, we examine the two flaw dimensions that are most important from a structural integrity point of view, flaw depth (dx) and flaw length (dy). Plate No. 1 was not used in the sizing analysis because of an extremely large flaw that caused problems in defining legitimate grading units; with Block 1, included in the analysis, some of the flaws were oversized by as much as 100 cm. This occurred when a small flaw was apparently too close to the large flaw in Block 1. An indication from the large flaw intersected with the grading unit of the small flaw and was associated with it, causing the problems.

4.1 Regression Models and Statistics for Evaluating Sizing

Several different regression models have been used to relate the measured and true dimensions of flaws. Davies (1987) has made the most comprehensive study of the various types of regression models that might describe the sizing process. His studies have concentrated on determining the distribution of the sizing error as well as the functional relationship between true and measured dimensions. Davies found that a log transformation of the sizes seemed to stabilize and symmetrize the errors. Therefore, he used a regression model of the form:

$$\begin{aligned}\log(M_i) &= \beta_1 + \beta_2 \log(T_i) + \epsilon_i \\ \text{Var}(\epsilon_i) &= \sigma_\epsilon^2 \\ \beta_2 &= 1\end{aligned}\tag{3}$$

where M_i is a measured flaw dimension (i.e., dy or dz) and T_i is the "true" flaw dimension. The errors ϵ_i are assumed to be approximately normally distributed with a constant variance σ_ϵ . Such a model assumes the errors are proportional to flaw size. This is in contrast to the most commonly used regression model for sizing (Heasler, et al. 1990), which does not use the log of the sizes and has the form

$$\begin{aligned}M_i &= \beta_1 + \beta_2 T_i + \epsilon_i \\ \text{Var}(\epsilon_i) &= \sigma_\epsilon^2\end{aligned}\tag{4}$$

In this case, the error made in sizing a small flaw is considered to be the same as that made in sizing a large flaw. The models also differ in the nature of the regression curve that is fitted to the sizing data. The logged model requires the regression curve to go through the origin, while the unlogged model does not restrict the curve to the origin.

Which regression model is correct? In our experience, we have found neither to be completely adequate, particularly in their representation of the sizing error structure. Neither model correctly accounts for the skewed nature of sizing errors (i.e., for small flaws, the errors are likely to be positive, and for large flaws, the errors negative).

In the case of the logged model, wave theory would argue against the proportional error structure, at least for small flaws. When flaws reach the dimension of the wave length being employed, the error cannot be reduced beyond the size of this wave-length. In fact, certain data (Heasler, et al. 1990) suggests that the sizing error for small flaws might actually INCREASE, as inspectors mistake random ultrasonic signals for the flaw edges. On the other hand, the errors made in sizing an extremely large flaw may be larger than sizing a medium sized flaw, simply because the reflecting edges are more discontinuous and variable. Factors

Sizing Performance

such as this would argue for some sort of proportional errors model.

To see which error structure most closely approximates that actually present in the PISC-II data, we calculated standard deviation as a function of crack size. The PISC-II data set allows such a calculation to be made because each flaw is measured several times by different teams. Thus a "replicate" error can be defined by:

$$M_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad (5)$$

where the index i represents a flaw, j a procedure type, and k the team "replicate". The standard deviation σ_{ij} of each flaw/procedure combination, can be estimated from the residuals ϵ_{ijk} in the standard manner. These standard deviations then can be plotted against true flaw size as illustrated in Figures 4.1 and 4.2. The un-logged model requires the standard deviation to be unrelated to flaw-size, while the logged model requires a linear relationship that goes through the origin. The standard deviations have been fit to a curve to allow the general trends in the data to be discerned.

Figure 4.1. Standard Deviation of Measured Depth Versus True Depth

Figure 4.2. Standard Deviation of Measured Length Versus True Length

As one can see from the plots, neither the depth nor length standard deviation seems to go to 0 as the respective size goes to 0, an indication that the logged-model is not valid for small flaw sizes. On the other hand, there does seem to be a modest increase in standard deviation with flaw size (in both the depth and length metrics), indicating that some form of variance stabilizing model might be justified.

Since the standard deviation is relatively constant throughout the range of sizes, the un-logged model appears most reasonable for this data set. The un-logged model is not sensitive to modest departures from the constant variance assumption. Therefore, we will employ the simple linear model for the present sizing analysis. This also has advantages because it will allow the present analyses to be compared with past work.

To evaluate a set of sizing results using the regression model (Eq. 4), one must compare the three regression parameters β_1 , β_2 , and σ_ϵ to their ideals. If the sizing procedure had done a perfect job, then $\beta_1 = 0$, $\beta_2 = 1$, and $\sigma_\epsilon = 0$; this is an indirect way of saying that all measured sizes equal the true sizes). Poor sizing will

be represented in a deviation of one or all of the regression parameters from their respective ideals. When β_1 , β_2 deviate from their ideals, a systematic bias is introduced into the measured sizes; and when σ_ϵ deviates from zero, a random error is introduced.

The PISC-II data set provides sizing information on literally hundreds of different procedure/ flaw/ material combinations. Sizing within a single combination can be fully analyzed by all the associated regression diagnostic statistics available (R-squared, t-statistics, influence, outliers, confidence bounds, etc.). However in order to be able to compare such a large number of results, we require the regression fits to be summarized in one or two statistics, which can be readily compared. The basic objective of a summary statistic should be quite simple; it should tell us how close the measured values are to the true values.

To be able to compare two different regression fits, and to order a set of fits from worst to best, we utilize the root mean square error (RMSE), a statistic that summarizes the three deviations of the regression parameters from their respective ideals. The root mean square statistic can be directly defined in terms of the original sizing data (M_i, T_i), $i=1 \dots n$ by the formula:

$$RMSE^2 = \frac{\sum_i (M_i - T_i)^2}{n} \quad (6)$$

Although it is not immediately obvious that this quantity has anything to do with the three regression parameters discussed above, a little algebra can show that

$$RMSE^2 = (\beta_1 + (\beta_2 - 1)\mu_T)^2 + (\beta_2 - 1)^2 \sigma_T^2 + \sigma_\epsilon^2 \quad (7)$$

$$\text{where } \mu_T = \frac{\sum T_i}{n}$$

$$\sigma_T^2 = \frac{\sum (T_i - \mu_T)^2}{n}$$

From this formula, it is evident that RMSE can only be zero when ideal regression results are obtained; in other words, $\beta_1 = 0$, $\beta_2 = 1$, and $\sigma_\epsilon = 0$. A small root mean square error forces the sizing bias (the first two terms in the equation) and the random error (the last term) to be small. Therefore RMSE summarizes the deviations of the regression parameters from the ideal. In this report, we will use the RMSE statistic as the principal means of comparing two different sizing regression fits. The sizing fit with the lowest RMSE value is the best.

Other statistics have also been used to summarize sizing performance. For example, the ASME Code, Section XI, Appendix VIII sizing test evaluates sizing results using the regression slope (β_2), the regression R-squared statistic (i.e., correlation), the maximum sizing error, and mean absolute deviation (MAD). Our evaluation of sizing results also include the MAD statistics, which is defined by:

$$MAD = \frac{\sum_i |M_i - T_i|}{n} \quad (8)$$

and may also be used to order the sizing results. In general, RMSE is more sensitive to large errors than MAD (a feature of RMSE that makes it a more stringent test for good sizing). It should also be noted that a formula exists that relates MAD to the regression parameters β_1 , β_2 , and σ_ϵ , but is more complicated than Eq. (8) and requires that the distributional form of the errors be known.

4.2 Comparison of U.S. with Non-U.S. Sizing Capability

The participating U.S. teams employed two major types of sizing procedures in PISC II, as illustrated in Table 4.1. Most of the sizing inspections performed by U.S. teams utilized amplitude drop procedures (20% and 50%), while a few employed diffraction techniques.

Table 4.1. Summary of Sizing Inspections by Procedure and Country

Procedure	Country	
	Non-U.S.	U.S.
20% DAC Amplitude Drop	48	16
50% DAC Amplitude Drop	38	1
Amplitude Drop	56	8
Combined	2	0
Diffraction	21	3
Physically Focused Technique	8	0
Computer Imaging	29	0
Simple Imaging	5	0
Total	186	28

Given this distribution of data, the most extensive comparison of U.S. with non-U.S. performance can be performed between the 20% amplitude drop procedures. Table 4.2 provides the essential depth regression fit statistics for the various flaw type and flaw location combinations present in the data set.

From Table 4.2, one can see that the RMSE statistic for depth sizing is approximately 2 cm. The highest RMSE in the table (3.66 cm) occurs for the U.S. procedures in base-metal when inspecting volumetric flaws. These volumetric depth measurements for U.S. teams exhibit large bias; the regression slope is negative, and the random error is also large. Another notably large RMSE is associated with U.S. measurements of smooth flaws in base-metal.

The overall RMSE is 2.37 cm for U.S. teams and 1.95 for Non-U.S. teams, not a large, but a significant difference. However, this comparison averages together

the high U.S. RMSE statistics from base metal smooth and volumetric flaws with much more reasonable performance. The overall RMSE obscures a very important relationship between the two categories of teams. It is important to note that all the U.S. RMSE statistics are larger than the non-U.S. in base-metal, while the situation is reversed for clad.

It so happens that the differences between U.S. and non-U.S. depth sizing performance were most severe for the 20% amplitude drop procedures. Table 4.3 presents averaged RMSE statistic comparisons for the 50% amplitude drop procedures and diffraction procedures (averaged over flaw type and location). In both these cases, the U.S. teams do slightly (but not significantly) better than the non-U.S. teams.

Table 4.4 presents the length regression statistics for 20% amplitude drop teams. The length results are generally consistent with the depth RMS statistics presented previously in Table 4.2. In this table, we also find the U.S. teams out-performing the Non-U.S. teams in clad and under-performing in base-metal, particularly for volumetric and smooth cracks. The results displayed by U.S. teams for volumetric flaws are particularly bad, with a RMSE of 8.7 cm. The volumetric flaws inspected by the U.S. teams were relatively large, with an average length of 8.82 cm and the flaws were widely distributed (standard deviation of 7.65 cm). It is not possible to ascribe these results to flaws that were too small to be accurately sized.

Overall, the U.S. teams obtained a RMSE of 4.09 (cm) while the non-U.S. teams did significantly better with a RMS error of 3.14. The reason for the inflated U.S. RMS error is, of course, the poor performance with smooth and volumetric flaws. If these two cases were eliminated, U.S. performance would be comparable to non-U.S. performance. For sizing in clad, we find that U.S. teams RMSE are generally lower than the respective non-U.S. teams' RMS as in the case of the depth RMSE's.

For the 50% amplitude drop procedures, the U.S. team's RMSE is actually better than non-U.S. teams, as illustrated in Table 4.5. In Table 4.5, the RMSE has been divided into the amount contributed by bias and random error (σ_e). From the table, one can see that bias contributes less to RMSE than random error, for

**Table 4.2. Comparison of U.S./Non-U.S. 20%-DAC Amplitude Drop
Depth Sizing Procedures**

Flaw Type/ Location	Country	Regression Statistics							
		NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Base-Metal									
Rough	U.S.	85	2.286	1.377	1.3432	0.890	1.644	1.963	1.484
Rough	Non-U.S.	191	2.610	1.599	0.941	0.5965	1.59	1.72	1.234
Smooth	U.S.	27	1.924	1.750	2.8120	0.546	2.546	3.220	2.241
Smooth	Non-U.S.	38	2.278	2.001	1.510	0.6475	2.39	2.53	1.942
Volumetric	U.S.	26	0.463	0.255	3.0694	-0.173	2.735	3.656	2.627
Volumetric	Non-U.S.	39	0.572	0.320	1.104	1.3617	1.45	1.93	1.446
Clad									
Rough	U.S.	20	3.280	1.624	1.3457	0.400	1.631	1.919	1.575
Rough	Non-U.S.	53	4.159	1.294	3.124	0.0591	1.74	2.23	1.847
Smooth	U.S.	35	2.534	1.392	1.4975	0.282	0.541	1.821	1.456
Smooth	Non-U.S.	43	3.391	1.509	1.918	0.3983	1.92	2.08	1.518
Volumetric	U.S.	14	0.236	0.173	0.3458	1.784	0.843	0.953	0.682
Volumetric	Non-U.S.	6	0.258	0.166	0.191	5.5429	1.12	1.78	1.365
Total	U.S.	207	1.787	1.267	1.4768	0.000	1.933	2.432	1.678
Total	Non-U.S.	370	2.211	1.332	1.098	0.000	1.75	2.06	1.559

Table 4.3. Summary of U.S./Non-U.S. Depth Sizing Performance

Procedure	Country	Regression Statistics			
		NOBS	Bias	SD(e)	RMSE
20%-DAC Amplitude Drop	U.S.	207	1.4856	1.896	2.365
	Non-U.S.	370	0.967	1.73	1.96
50%-DAC Amplitude Drop	U.S.	101	1.8128	1.739	2.440
	Non-U.S.	864	1.348	2.24	2.60
Diffraction	U.S.	20	1.1381	0.989	1.407
	Non-U.S.	181	0.949	1.37	1.63

all three types of procedures listed and for both U.S. and non-U.S. teams.

The sizing results can be graphically compared also. Figures 4.3 through 4.10 present comparisons of U.S. versus non-U.S. inspections for rough and smooth flaws. Perhaps the most significant general feature of these plots is the substantial flaw-to-flaw variability that they display; each vertical row of points in the plots corresponds to measurements taken on a single flaw. From the plots, one can see that some flaws' sizes are substantially above the regression line, and others are below it. An extreme example of this is displayed in Figure 4.9 which displays the poor results obtained by U.S. teams on smooth, base-metal flaws. From this plot, one can see that the poor results are almost entirely due to one small flaw, which was oversized by a gross amount.

This RMSE comparison of U.S. versus non-U.S. teams shows that both U.S. and non-U.S. teams sizing capability are comparable, except for sizing smooth and volumetric flaws in base metal. For these types of flaws, U.S. teams did not do as well. Appendix D contains a more complete set of sizing regression curves of depth and length for Advanced, Manual-10% DAC, Manual-20% DAC, and Manual-50% DAC procedures.

4.3 Comparison of Clad Versus Base-Metal Sizing Capability

Tables 4.6 and 4.7 present RMSE statistics that can be used to compare sizing performance for clad versus the base-metal flaws. To compute these statistics, all teams have been used (U.S. and non-U.S.). The tables display much variation in the RMSE statistic; for some procedures, performance in clad is substantially better than base-metal. In other procedures, the relationship is reversed.

One would expect to see a consistency within the amplitude drop procedures, if flaws in one location were easier to size than another. In the tables, clad is better 8 out of 14, an outcome that could be ascribed to chance.

4.4 Comparison of Sizing Procedures

Tables 4.8 and 4.9 provide a comparison of the various types of sizing procedures present in the PISC-II data set. In these tables, two procedures have been deleted, because of insufficient data. These procedures are the combination procedures and the simple imaging procedures.

**Table 4.4. Comparison of U.S./Non-U.S. 20%-DAC Amplitude Drop
Length Sizing Procedures**

Flaw Type/ Location	Country	Regression Statistics							
		NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Base-Metal									
Rough	U.S.	85	6.86	1.89	1.711	0.72405	3.90	3.90	3.196
Rough	Non-U.S.	191	7.15	1.75	-0.7599	0.8303	2.74	3.38	2.64
Smooth	U.S.	27	4.39	1.87	8.828	-0.45881	5.35	6.29	3.267
Smooth	Non-U.S.	38	4.34	1.73	0.4375	0.8987	2.26	2.21	1.92
Volumetric	U.S.	26	8.82	7.65	4.863	0.21510	6.33	8.71	6.208
Volumetric	Non-U.S.	39	7.84	5.64	-1.1736	1.0665	3.12	3.13	2.23
Clad									
Rough	U.S.	20	3.71	1.76	2.057	0.53242	1.81	1.92	1.109
Rough	Non-U.S.	53	4.75	1.01	3.4518	0.5920	2.48	2.89	2.22
Smooth	U.S.	35	4.04	2.85	2.206	0.26419	2.32	3.15	2.314
Smooth	Non-U.S.	43	3.02	2.45	1.7223	0.8954	2.75	3.04	2.08
Volumetric	U.S.	14	4.46	2.27	1.374	0.55043	2.75	2.80	2.188
Volumetric	Non-U.S.	6	4.54	2.16	5.9169	0.0831	2.13	3.06	2.40
Total	U.S.	207	5.38	3.70	2.943	0.000	4.09	5.03	3.047
Total	Non-U.S.	370	5.27	2.87	1.4416	0.000	2.60	2.98	2.25

Table 4.5. Summary of U.S./Non-U.S. Length Sizing Performance

Procedure	Country	Regression Statistics			
		NOBS	Bias	σ_{ϵ}	RMSE
20%-DAC Amplitude Drop	U.S.	207	2.775	4.09	4.85
	Non-U.S.	370	1.6761	2.70	3.14
50%-DAC Amplitude Drop	U.S.	101	2.338	3.72	4.23
	Non-U.S.	864	2.0563	4.67	5.08
Diffraction	U.S.	20	3.361	3.04	4.21
	Non-U.S.	181	2.2575	3.40	3.99

Of the four flaw type/locations listed in the tables, we find that in all but one case (length sizing for rough cracks in clad), the advanced procedures exhibit the best RMSE. In most cases, the advanced procedures exhibit RMSE statistics that are substantially better than the amplitude drop procedures. Within the advanced procedures, no clear winner exists; each of the three advanced procedures obtains a low RMSE for some of the flaw types and a high RMSE for others.

4.5 Summary

A simplified set of statistics was developed for the PISC-II sizing data that has permitted conclusions to be drawn from this data base. The procedure used most extensively by U.S. teams was a 20% amplitude drop-based procedure. Extensive analysis was performed on this data to determine if there was a difference in performance of the U.S. versus the non-U.S. teams. It was determined that even though minor differences existed, the data could be collapsed for both U.S. and non-U.S. teams for trend analysis.

When all of the data for U.S. and non-U.S. teams was combined and a comparison was made between the performance in clad versus the base-metal, the conclusion was that there was no systematic trend in the data and that one location was not easier to size flaws in than the other.

The effectiveness of the procedures was examined in great detail, and the advanced procedures were superior to the amplitude drop procedures. Within the advanced procedures, there was no clear winner because each was superior for some flaw type/location conditions and worse for others.

Table 4.6. Summary of Clad Versus Base-Metal Depth Sizing Performance

Procedure	Flaw Location	Regression Statistics			
		NOBS	Bias	σ_{ϵ}	RMSE
20% DAC Amplitude Drop	Base	406	1.34	2.12	2.51
	Clad	171	1.00	1.54	1.83
50% DAC Amplitude Drop	Base	226	1.46	1.75	2.28
	Clad	106	2.28	2.82	3.63
Amplitude Drop	Base	398	1.55	2.30	2.77
	Clad	235	1.12	2.10	2.38
Total	Base	1030	1.45	2.07	2.53
	Clad	512	1.58	2.22	2.72
Diffraction	Base	137	0.835	1.473	1.69
	Clad	68	0.932	1.237	1.55
Physically Focused Technique	Base	51	0.943	1.431	1.71
	Clad	10	2.528	1.000	2.72
Computer Imaging	Base	142	0.879	1.778	1.98
	Clad	71	1.196	0.999	1.56
Total	Base	330	0.887	1.57	1.80
	Clad	150	1.568	1.09	1.91

Table 4.7. Summary of Clad Versus Base-Metal Length Sizing Performance

Procedure	Flaw Location	Regression Statistics			
		NOBS	Bias	σ_{ϵ}	RMSE
20% DAC Amplitude Drop	Base	406	1.96	4.34	2.89
	Clad	171	1.12	2.66	2.89
50% DAC Amplitude Drop	Base	226	1.44	4.16	4.40
	Clad	106	3.45	5.57	6.55
Amplitude Drop	Base	398	1.65	4.19	4.50
	Clad	235	3.06	6.32	7.02
Total	Base	1030	1.70	4.23	4.56
	Clad	512	2.74	5.1	5.79
Diffraction	Base	137	1.950	3.86	4.33
	Clad	68	1.752	2.90	3.39
Physically Focused Technique	Base	51	NA	4.44	4.38
	Clad	10	0.567	2.80	2.86
Computer Imaging	Base	142	1.365	3.80	4.06
	Clad	71	1.097	2.18	2.44
Total	Base	330	1.3	4.05	4.26
	Clad	150	1.3	2.63	2.93

Table 4.8. Comparison of Depth Sizing Regressions

Location/ Type Procedure	Regression Statistics							
	NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Base/Rough								
20% DAC Amplitude Drop	276	2.510	1.539	1.170	0.6369	1.693	1.795	1.31
50% DAC Amplitude Drop	170	2.626	1.558	2.243	0.4243	1.826	2.152	1.61
Amplitude Drop	229	2.615	1.609	2.031	0.5329	2.355	2.591	1.75
Diffraction	88	2.540	1.586	1.037	0.5311	1.278	1.472	1.05
Physically Focused Technique	27	2.193	1.478	1.502	0.3998	1.319	1.550	1.03
Computer Imaging	87	2.784	1.742	1.683	0.5328	1.306	1.571	1.14
Base/Smooth								
20% DAC Amplitude Drop	65	2.131	1.894	2.102	0.5855	2.480	2.838	2.06
50% DAC Amplitude Drop	23	1.820	1.678	1.768	0.8508	1.759	2.263	1.72
Amplitude Drop	78	2.522	2.243	1.926	1.0236	2.549	3.205	2.32
Diffraction	22	3.211	2.738	1.087	0.5865	1.844	2.091	1.30
Physically Focused Technique	12	2.946	2.539	1.269	0.5812	1.741	1.888	1.40
Computer Imaging	19	2.982	2.639	1.316	0.6452	2.475	2.526	1.67
Clad/Rough								
20% DAC Amplitude Drop	73	3.918	1.436	2.270	0.2310	1.717	2.150	1.77
50% DAC Amplitude Drop	43	4.080	1.324	3.778	-0.0641	2.434	2.812	2.37
Amplitude Drop	84	3.712	1.472	2.464	0.3824	1.933	2.119	1.73
Diffraction	24	4.087	1.376	2.036	0.2765	1.063	1.683	1.12
Physically Focused Technique	5	3.940	1.360	4.634	-0.7650	1.086	3.271	2.68
Computer Imaging	32	4.267	1.151	3.085	0.0220	1.020	1.841	1.20
Clad/Smooth								
20% DAC Amplitude Drop	78	3.006	1.510	1.526	0.4219	1.776	1.968	1.49
50% DAC Amplitude Drop	42	3.569	1.429	3.804	-0.2512	2.264	2.905	2.38
Amplitude Drop	117	2.946	1.455	1.911	0.4620	2.250	2.386	1.63

Location/ Type Procedure	Regression Statistics							
	NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Diffraction	34	3.159	1.512	1.032	0.4825	1.523	1.772	1.37
Physically Focused Technique	5	2.840	1.815	1.792	0.0132	0.905	2.019	1.65
Computer Imaging	32	3.897	1.244	2.887	0.0440	1.248	1.879	1.61

Table 4.9. Comparison of Length Sizing Regressions

Location/ Type Procedure	Regression Statistics							
	NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Base/Rough								
20% DAC Amplitude Drop	276	7.06	1.80	0.267	0.7608	3.230	3.544	2.80
50% DAC Amplitude Drop	170	7.21	1.71	5.710	0.1917	5.204	5.356	3.29
Amplitude Drop	229	6.95	2.07	1.633	0.6628	4.116	4.217	3.05
Diffraction	88	7.06	2.02	1.437	0.4586	3.370	4.238	3.15
Physically Focused Technique	27	6.50	2.15	2.603	0.7595	5.845	5.742	3.82
Computer Imaging	87	7.17	1.93	2.706	0.6147	3.716	3.747	2.74
Base/Smooth								
20% DAC Amplitude Drop	65	4.36	1.77	4.066	0.2984	4.154	4.389	2.48
50% DAC Amplitude Drop	23	4.20	1.93	-0.170	1.2924	2.092	2.328	1.76
Amplitude Drop	78	4.85	2.15	5.028	0.2041	4.588	4.975	2.77
Diffraction	22	5.16	2.39	4.560	0.0421	4.565	4.907	3.07
Physically Focused Technique	12	5.23	2.49	3.209	0.3350	1.533	2.134	1.69
Computer Imaging	19	4.93	2.45	2.463	0.5173	3.150	3.195	2.50
Clad/Rough								
20% DAC Amplitude Drop	73	4.46	1.331	2.277	0.756	2.396	2.664	1.91
50% DAC Amplitude Drop	43	4.71	1.092	10.313	-0.972	2.923	3.706	2.33
Amplitude Drop	84	4.31	1.484	6.559	-0.229	3.316	3.953	2.45

Location/ Type Procedure	Regression Statistics							
	NOBS	μ_T	σ_T	β_1	β_2	σ_ϵ	RMSE	MAD
Diffraction	24	4.62	1.146	1.237	0.299	2.607	3.297	2.71
Physically Focused Technique	5	4.52	1.236	6.937	-0.530	3.902	3.463	3.10
Computer Imaging	32	4.93	0.654	-3.025	1.721	2.789	2.791	2.14
Clad/Smooth								
20% DAC Amplitude Drop	78	3.48	2.667	2.616	0.503	2.801	3.093	2.18
50% DAC Amplitude Drop	42	2.67	2.263	3.504	0.692	5.174	5.760	3.07
Amplitude Drop	117	3.70	2.429	2.662	0.462	3.330	3.612	4.75
Diffraction	34	3.67	2.220	1.183	0.525	3.084	3.216	2.32
Physically Focused Technique	5	4.34	1.704	0.514	0.458	0.702	2.090	1.84
Computer Imaging	32	2.96	2.527	1.852	0.668	1.614	1.970	1.68

Figure 4.3. U.S. 20%-DAC Amplitude Drop Depth Measurements of Rough Flaws in Base Metal

Figure 4.4. Non-U.S. 20%-DAC Amplitude Drop Depth Measurements of Rough Flaws in Base Metal

Figure 4.5. U.S. 20%-DAC Amplitude Drop Length Measurements of Rough Flaws in Base Metal

Figure 4.7. U.S. 20%-DAC Amplitude Drop Depth Measurements of Smooth Flaws in Base Metal

Figure 4.6. Non-U.S. 20%-DAC Amplitude Drop Length Measurements of Rough Flaws in Base Metal

Figure 4.8. Non-U.S. 20%-DAC Amplitude Drop Depth Measurements of Smooth Flaws in Base Metal

Figure 4.9. U.S. 20%-DAC Amplitude Drop Length Measurements of Smooth Flaws in Base Metal

Figure 4.10. Non-U.S. 20%-DAC Amplitude Drop Length Measurements of Smooth Flaws in Base Metal

5.0 Use of ASME Criteria to Evaluate PISC-II Procedures

Section XI of the ASME Code recently adopted Appendix VIII, "Performance Demonstration for Ultrasonic Examination Systems." This appendix defines requirements for performance demonstration of ultrasonic inspection systems used for inservice inspection of piping and reactor vessels. This section of the report assesses the detection and sizing performance of teams that participated in the PISC-II round robins with regard to Appendix VIII performance demonstration requirements.

Appendix VIII contains performance demonstration requirements for the following areas of the reactor pressure vessel.

1. Clad/Base Metal Interface Region
2. Vessel Shell Welds
3. Nozzle-to-Vessel Welds
4. Nozzle Inside Radius Section

Of the four parts of the ASME test listed above, only the first three were applicable to the PISC-II data; those dealing with the inside radius examinations could not be evaluated because the results of inner radius examinations were not included in the PISC data file "CLEAN.PROC". The design of the PISC-II test specimens did not allow a sufficient number of blank grading units to test the false call portion of the Appendix VIII performance demonstration. Also, the PISC-II data was "cleaned" and most false call indications were eliminated. Consequently, the passing or failing of a performance demonstration test structured from the PISC-II data was essentially only based on the flaw detection criteria. The acceptance criteria outlined in Appendix VIII for detection (sans false call criteria) were applied to the PISC-II "summary" inspections associated with each team. The sizing criteria were also applied to the summary inspections to analyze PISC-II sizing capabilities with respect to the new ASME Code requirements.

The specific specimen, testing, and acceptance requirements for Appendix VIII performance demonstrations are provided in Appendix D. The

specific test from Appendix D that was used for detection was an 8 of 8 test. The performance demonstrations outlined by ASME Code are based on the concept that inspection performance can be quantified by estimating probability of detection and false call probability. The performance demonstrations in Appendix VIII estimate the probability of detection and false call probability, and the acceptance criteria associated with the tests screen inspection systems which poor detection and sizing capability. The performance demonstrations in Appendix VIII are designed such that inspections that have a false call probability greater than 20% and a probability of detection less than 70% have less than a 1% chance (probability) of success. Alternately, inspections with a probability of detection of 90% or greater and a false call probability less than 20% will have an 80% percent or greater chance of success.

One notices that the acceptance criteria in Appendix D are more stringent for smaller sample sizes. This restriction on small sample sizes accounts for experimental error in the tests.

In order to apply the ASME test to the PISC-II data, a subset of flaws must be chosen that satisfies the ASME requirements. We selected a subset of 20 flaws from Plate 2, Nozzle 3, and Nozzle 9 to analyze PISC-II performance for detection of defects at the clad/base metal region. Plate #1 was not used in this analysis because it contained a very large flaw that biased sizing results. The large flaw did not allow construction of a grading unit (see Section 3.1) that would provide unbiased sizing scoring. Table 5.1 below provides summary statistics for the flaws used to analyze the detection and sizing performance of PISC-II participants relative to an Appendix VIII performance demonstration for flaws at the clad/base metal region.

Table 5.2 provides a similar summary of the flaws used to analyze the performance of PISC-II participants in detecting and sizing flaws in reactor vessel shell welds other than the clad/base-metal region. Again, Plate #1 was not used in this analysis for the same reasons as described above.

Table 5.1. Summary of Flaws Used in ASME-Type Clad Test

Block Type		Flaw Type			
		Multiple	Rough	Smooth	Volumetric
NOZZLE Depth, cm	No. of Flaws	NA	2	9	1
	Mean	NA	1.75	2.36	0.60
	Minimum	NA	1.00	1.00	0.60
	Maximum	NA	2.5	5.0	0.6
PLATE Depth, cm	No. of Flaws	2	3	3	NA
	Mean	2.68	4.63	4.67	NA
	Minimum	1.05	3.55	4.50	NA
	Maximum	4.3	6.1	4.9	NA

Table 5.2. Summary of Embedded Flaws Used in ASME-Type Shell Weld Detection Test

Block Type		Flaw Type			
		Multiple	Rough	Smooth	Volumetric
NOZZLE Depth, cm	No. of Flaws	1	12	9	11
	Mean	0.60	1.82	2.94	0.32
	Minimum	0.60	0.3	1.00	0.20
	Maximum	0.6	3.4	8.0	0.50
PLATE Depth, cm	No. of Flaws	NA	5	NA	2
	Mean	NA	4.06	NA	0.73
	Minimum	NA	1.00	NA	0.35
	Maximum	NA	6.50	NA	1.10

5.1 ASME Performance Demonstration Results for Clad/Base Metal Interface Region

The clad/base metal region of the reactor vessel includes the region of the clad/base metal interface plus the first quarter wall thickness. This region of the vessel is inspected to detect defects that might become important, for instance, in a pressurized thermal shock event. In this section, the ability of inspections performed during the PISC II round robins to pass an Appendix VIII performance demonstration-type test for detection and sizing in the clad/base metal interface region is evaluated for the following cases:

1. Flat plate (Test block 2) examined from the clad side
2. Flat plate examined from the unclad side
3. Nozzle components (Nozzle 3 and 9) examined from the clad side
4. Nozzle components examined from the unclad side

5.1.1 Detection Results for the Clad/Base Metal Region

Table 5.3 summarizes the results of an ASME-type clad test (e.g., using new ASME Section XI, Appendix VIII

Table 5.3. Summary of Teams Passing ASME-Type Clad Detection Test

Procedure		Nozzle		Plate		Total
		Clad	Unclad	Clad	Unclad	
Advanced	No. Passed	1	0	3	2	6
	No. Tested	3	1	3	6	13
Automated-10% DAC	No. Passed	0	0	1	NA	1
	No. Tested	1	1	1	NA	3
Automated-20% DAC	No. Passed	0	0	NA	0	0
	No. Tested	1	1	NA	2	4
Automated-50% DAC	No. Passed	NA	0	0	0	0
	No. Tested	NA	1	2	3	6
Manual-10% DAC	No. Passed	0	0	1	1	2
	No. Tested	1	NA	2	2	4
Manual-20% DAC	No. Passed	1	1	6	2	10
	No. Tested	4	7	8	5	24
Manual-50% DAC	No. Passed	NA	0	0	1	1
	No. Tested	NA	2	1	5	8
Total	No. Passed	2	1	11	6	20
	No. Tested	10	13	17	23	63

Table 5.4. Summary of Teams Passing ASME-Type Clad Sizing Test

Access/ Block Type	Sizing Procedure						
	Advanced	Auto ¹ 10%-DAC	Auto 20%-DAC	Auto 50%-DAC	Manual 10%-DAC	Manual 20%-DAC	Manual 50%-DAC
<i>Clad/Nozzle</i>							
Pass-Depth	0	0	0	0	0	0	0
Pass-Length	0	0	0	0	0	1	0
Pass-Both	0	0	0	0	0	0	0
No. of Teams	2	0	0	0	0	3	0
<i>Unclad/Nozzle</i>							
Pass-Depth	0	0	0	0	0	0	0
Pass-Length	0	0	0	0	0	0	0
Pass-Both	0	0	0	0	0	0	0
No. of Teams	0	1	0	0	0	1	0
<i>Clad/Plate</i>							
Pass-Depth	0	0	0	0	0	0	0
Pass-Length	0	0	0	0	0	1	0
Pass-Both	0	0	0	0	0	0	0
No. of Teams	2	1	0	0	0	6	0
<i>Unclad/Plate</i>							
Pass-Depth	0	0	0	0	0	0	0
Pass-Length	0	0	0	0	0	0	0
Pass-Both	0	0	0	0	0	0	0
No. of Teams	3	0	0	0	1	2	1

¹Auto = Automated

criteria) on PISC-II teams. Of the 63 team/procedure combinations that this test was applied to, 32% (20 team/procedures) passed the test, while the rest failed. One of the reasons for poor performance is related to the procedures the teams used. Few teams used

techniques specifically designed to detect near-surface defects. Of the 27 teams in Table 5.3 that examined the test blocks from the clad side, only 15 teams used inspection techniques designed to detect near-surface flaws. Nine of these 15 teams passed the test, while

only four of 12 teams that did not use a special near-surface technique passed.

The message from PISC-II data is quite clear: inspection techniques designed for detecting near-surface flaws must be used to pass an ASME performance demonstration test.

The Manual-20% group of teams seems to have produced the highest proportion of passes overall (inspections from clad and unclad access), with 42% of the teams passing overall and 75% passing the test on Plate #2 from the clad side. Advanced procedures also performed very well with 3 out of 3 teams passing the test on Plate #2 from the clad side.

5.1.2 Sizing Results for the Clad/Base Metal Region

The same set of flaws used to analyze detection performance were used to analyze sizing performance relative to the new ASME sizing performance criteria. The subset of flaws used for sizing did not contain the required minimum of ten flaws; however, we conducted the analysis anyway and only eight flaws were used in the analysis. Table 5.4 summarizes the results of the sizing test. No teams passed the depth sizing test and only one team passed the length sizing test. The performance criteria most often the cause for failure was "mean deviation." The results of comparing PISC-II data with ASME sizing criteria is consistent with the results of Section 4.0.

5.2 Detection Results for ASME-Type Vessel Weld Test

The results from analyzing the PISC-II data using performance demonstration criteria as outlined by ASME Appendix VIII are summarized in Table 5.5. For the vessel shell criteria (tested by examination of Plate #2), 45% of the teams passed (18 of 40), while a much smaller proportion (10%) (6 of 58) passed the nozzle-to-vessel (tested by examination of Nozzles 3 and 9) criteria. The Manual-10% DAC teams scored best with 4 of 4 teams passing the detection test on the plate test block.

If one compares the overall pass rate for the inspection variable "access" (e.g., whether the inspection was performed from the clad or unclad side of the test specimen), the detection data in Table 5.5 indicates that inspection from the "clad" side is better than inspection from the "unclad" side. Upon closer inspection, however, the inspection data from the unclad side contains many team/procedures that used a 50% DAC recording level, which is not effective for defect detection. If one eliminates the team/procedures that used 50% DAC recording level, the data in Table 5.5 shows very little difference in detection capability between clad and unclad inspection access.

The poor results for the nozzle-to-vessel shell inspections were experienced by each type of procedure. The geometry of Nozzle 3 was complex; this may be the cause of the low scores.

5.2.1 ASME Vessel Weld Sizing Results

The results from analyzing PISC-II inspection for sizing defects contained within the vessel shell or nozzle are as discouraging as those of the clad/base metal region. Only two teams passed the depth sizing criteria for shell welds and no teams passed the depth criteria for nozzles. The performance criteria most often causing failure of depth sizing was "mean deviation." The results are consistent with Section 4.0.

5.3 Summary

The PISC-II detection inspection capability was evaluated using criteria found in the newly adopted ASME Code Section XI, Appendix VIII entitled "Performance Demonstration for Ultrasonic Examination Systems." These new ASME criteria were applied to the PISC-II "Summary" inspection data from each team. The new Section XI criteria are more stringent for smaller sample sizes; this is necessary in order to properly account for experimental error. Some of the Appendix VIII requirements could not be applied to the PISC-II data due to the limited amount of blank material present within the PISC-II blanks, and the "cleaned" data set.

Table 5.5. Summary of Passing Teams for ASME-Type Vessel Weld Detection Test

Procedure		Nozzle-Vessel			Vessel-Shell		
		Clad	Unclad	Total	Clad	Unclad	Total
Advanced	No. Passed	2	0	2	1	3	4
	No. Tested	3	4	7	3	6	9
Automated-10% DAC	No. Passed	0	0	0	1	NA	1
	No. Tested	1	2	3	1	NA	1
Automated-20% DAC	No. Passed	1	0	1	NA	1	1
	No. Tested	2	2	4	NA	2	2
Automated-50% DAC	No. Passed	1	0	1	0	0	0
	No. Tested	2	2	4	2	3	5
Manual-10% DAC	No. Passed	0	1	1	2	2	4
	No. Tested	2	3	5	2	2	4
Manual-20% DAC	No. Passed	1	0	1	5	3	8
	No. Tested	9	15	24	8	5	13
Manual-50% DAC	No. Passed	0	0	0	0	0	0
	No. Tested	3	8	11	1	5	6
Total	No. Passed	5	1	6	9	9	18
	No. Tested	22	36	58	17	23	40

A subset of 20 flaws were selected from within the clad/interface region for evaluation using the ASME performance demonstration criteria. These flaws were all located within 1/4 wall thickness of the cladding for each plate. Of the 63 team/procedure combinations to which this test were applied, 32% would have passed the ASME-type performance demonstration criteria. One reason for such poor performance was that of the 27 teams that examined Plate #2, Nozzle #3 and Nozzle #9, only 15 used inspection techniques designed to detect near-surface flaws. Nine of the 15 passed, while only four of the 12 teams that did not use special techniques for the clad/base metal interface passed the test. Thus, it may be concluded that only

inspection techniques designed for detecting near-surface flaws should be used when attempting to detect flaws in the clad/base-metal region. The Manual-20% DAC teams exhibited the best performance overall (42% of these passed); which is consistent with other situations in which the Manual-20% DAC teams also performed well. It was also found that the performance of the clad-side inspections was much better than those performed from the unclad side when including teams using 50%-DAC procedures. If these 50%-DAC teams are eliminated, then there was the same level of performance for inspections conducted from both the clad and unclad inspection surfaces.

The PISC-II results for flaws located within the vessel shell weld metal were evaluated using the ASME Appendix VIII criteria for vessel shell examinations and nozzle-to-vessel welds. Forty flaws in PISC-II Plate #2, Nozzle #3 and Nozzle #9 were selected for this portion of the analysis. These results were somewhat more favorable than those from the clad/interface test described above. For the vessel shell data, 45% of the teams would have passed an ASME-type test based on the Appendix VIII criteria; and for the nozzle-to-vessel weld data, only 10% of the teams would have passed an ASME-type test based on the ASME criteria. The best results were achieved by the Manual-10% DAC teams who achieved an impressive 100% pass rate on the vessel shell examination. None of the techniques used were particularly effective for the nozzle-to-vessel weld examination; however, it should be noted that the nozzle geometry in Plate #3 was quite complex and presented a difficult inspection problem.

The results for sizing performance were quite poor for an ASME-type test based on Appendix VIII criteria. None of the teams passed the depth sizing criteria for cracks at the clad/base metal region, none passed the depth sizing criteria for flaws in the nozzle vessel weld, and only two teams passed the depth sizing test for vessel shell welds. The poor sizing results presented in this section are in contrast to results presented in "The PISC II Final Report" (Nichols and Crutzen, 1988). The difference between the results presented in this report and the results presented by Nichols and Crutzen (1988) are due, in part, to the analysis scheme used for this report. For this report, the largest indication intersecting the grading unit was used for sizing. The results presented by Nichols and Crutzen were analyzed to provide an upper bound on sizing capability.

6.0 Conclusions and Recommendations

The re-analysis of the PISC-II data base using more formal statistical methods was accomplished. In this re-analysis, the detection performance for U.S. teams was found not to differ greatly from similar teams from other countries. Sizing performance of U.S. teams is also generally equivalent to non-U.S. capabilities. There were some exceptions in which teams from the U.S. had lower performance; for instance, in the detection of volumetric flaws. However, it is reasonable to utilize all the PISC-II round robin test data to evaluate questions concerning the ultrasonic inspection capability of teams and procedures in the U.S.

6.1 Conclusions Regarding Probability of Detection

The probability of detection results reported in Section 3.0 are generally lower than those reported by PISC II in Report No. 5. The difference in results occurs because the analysis methods used by PISC II were different than those used in this report. The results reported by PISC II were achieved by extensive committee interpretation of results and by programming a computer to select the best results from each procedure or technique used by a team and then reporting the best combined result. The results reported by PISC II in Report No. 5 were intended to provide optimum inspection results. The probability of detection results reported in Section 3.0 were obtained by developing grading units that contained flawed and unflawed material and scoring the summary results presented by each team (e.g., data recorded on data sheets 6.2). The topic of data analysis and experimental design is discussed in greater detail below. In addition to the fundamental difference in analysis methodology, our report indicates:

1. The POD appears to be better for higher sensitivity procedures. Our analysis indicates that POD was highest for Manual-10% DAC, followed by Manual-20% DAC, and with Manual-50% DAC coming in last.
2. Flaws that are located in the clad/base metal region require a procedure designed specifically for the clad/base metal for adequate detection.

3. The simple logistic curve model for POD does not fit the data as theory would indicate; the variability of the data points is much more than that which should be exhibited in binomial data. The variations are associated with individual flaws; therefore, the detection procedures are influenced by flaw-to-flaw variability.

6.2 Conclusions Regarding Sizing

The results of sizing reported in Section 4.0 are considerably worse than those reported by PISC II in Report No. 5. The reason for the difference in results is due to analysis methodology. The sizing results reported by PISC II were the result of a committee (Data Analysis Group) reviewing inspection results and deciding which indications should be associated with any given flaw to determine flaw sizing. The flaw sizing results reported in Section 4.0 were based on using those indications determined by an algorithm which selected the largest indication reported by a team that intersected the flaw cuboid. The analysis of PISC-II sizing data reported in Section 4.0 also supports the following:

1. Advanced sizing procedures (e.g., SAFT, TOTF, ALOK, holography) seemed to characterize flaws (in both length and depth) better than other procedures used during the PISC-II round robin.
2. Flaw-to-flaw and team-to-team variability are important features in sizing measurements. This finding illustrates the need for serious study of the human element in NDE.
3. The error structure of flaw sizing is complicated and none of the sizing models proposed in Section 4.0 is completely adequate.

6.3 Conclusions Regarding PISC-II Data and ASME Section XI

Comparing the capability of ultrasonic inspection as measured by the PISC-II round robin exercise with

Conclusions and Recommendations

ASME Section XI performance criteria specified in Appendix VIII indicates:

1. The detection performance demonstration requirements proposed by ASME Section XI clearly screen inspection capability well.
2. Ultrasonic inspection technology clearly has the capability to meet the performance criteria specified by ASME Section XI for flaw detection in thick section steel and for complex geometries such as nozzles.
3. Procedures specifically designed for flaw detection at the clad/base metal interface have the highest probability for passing an ASME-type performance demonstration test for flaws at that location.
4. Ultrasonic inspection technology clearly needs to improve flaw length and sizing capability. No PISC-II teams were able to pass the sizing performance requirements outlined in Appendix VIII for the clad/base metal region and only two teams passed the depth sizing test for the vessel shell welds.
5. Large variations in performance from team-to-team using the same procedure/technique indicate that many individual teams/inspectors will fail ASME performance criteria until ultrasonic inspection processes are understood well enough to ensure consistently acceptable inspection practices.

6.4 Conclusions Regarding Experimental Design

During the analysis presented in this report, several problems occurred which merit discussion. First, attempting to control the analysis for specific variables was very difficult. In most instances, the difficulty was rooted in the lack of an experimental design. While the objectives of PISC II were clearly defined (Nichols and Crutzen, 1988), the experimental design used to gather data to achieve the goals was not well defined. The lack of a well-defined experimental design limits the

utility of the PISC-II data. Two rather simple guidelines could aid the design of future round robins. These are:

1. each (or nearly each) team examines the same set of flaws, and
2. each flaw is examined by the same set of teams.

These two requirements assure that teams (and flaws) can legitimately be compared against each other. We recognize that for an international program like PISC where the teams participate at their own expense, there may be difficulties in meeting these two requirements. However, an experimental design could be useful in making a case to the teams as to the significance of their contribution and the importance it plays in creating an adequate data base for reaching definitive conclusions.

If the characteristics of the inspections are to be varied within the round robin, it is important to plan these variations in a systematic way so that variables are not "confounded" with each other and there are sufficient observations to allow estimation of POD and sizing ability within each set of conditions.

A final criteria is critical in the design of a round robin -- blank material should be included in the study as a "control." In order to properly determine detection performance, one must be able to estimate probability of detection AND false call probability.

6.5 Conclusions Regarding Practice and Recommendations for Regulation of Ultrasonic Inspection of Reactor Pressure Vessels

This section provides a summary of conclusions regarding the practice of ultrasonic inspection of reactor vessels, a status of recent ASME Code requirements relative to each practice, and recommendations regarding regulation of ultrasonic inspection of reactor pressure vessels.

The following major conclusions regarding performance of ultrasonic inspection of reactor pressure vessels were presented in the Final PISC-II Report (Nichols and Crutzen, 1988) and are substantiated by this report. These conclusions are denoted below as (a), (b), (c), and (d) in a shortened format.

- (a) Decreasing the DAC level for an ASME type examination from 50% to 20% increases the detection rate.
- (b) Procedures in the spirit of ASME Section XI appear to perform better if a 70° longitudinal angle probe is added either in the pulse echo technique or with dual beam, or both. Tandem probes can also make an important contribution in some cases.

Recommendation: The 1986 Addenda to Section XI should be adopted by the NRC because it requires a 20% DAC recording level and requires the use of a 70° longitudinal angle probe as a result of these two conclusions.

- (c) Defect detection frequency appears to be independent of defect position through the plate thickness; however, sizing accuracy does appear to depend on position.

- (d) Defect size appears to affect both detection reliability and sizing accuracy.

The performance demonstration requirements specified in Appendix VIII and adopted in the 1989 Addenda to Section XI (summarized in Appendix D) require that defects be distributed throughout the vessel and that defects vary in size. However, the variation in flaw sizes specified in Appendix VIII is very limited. In fact, the flaws specified in Appendix VIII are so limited that the sizing performance test could be passed by analytical testmanship.

Recommendation: NRC staff should adopt Appendix VIII and carefully review any proposed test standards for Appendix VIII to ensure that an adequate variation exists to preclude testmanship.

7.0 References

- Bates, D. J., S. R. Doctor, P. G. Heasler, and E. Burck. 1987. Stainless Steel Round Robin Test: Centrifugally Cast Stainless Steel Screening Phase. NUREG/CR-4970, PNL-6266, PISC III Report No. 3. Prepared by Pacific Northwest Laboratory for the U.S. Nuclear Regulatory Commission, Washington, D.C.
- Crutzen, S. J. May 1979. PISC-I Reports, Plate Inspection Steering Committee, Joint Research Centre, Ispra, Italy.
- 1) A Description of the PISC Project
 - 2) Ultrasonic Examination of the PVRC Plates
 - 3) Destructive Examination of the PVRC Plates
 - 4) Analysis Scheme of the PISC Trials Results
 - 5) Evaluation of PISC Trials Results.
- Davies, J. K. W. 1987. A Statistical Model for the Analysis of Sizing Performance in the PISC-II RRT. Ispra Report PISCDOC(87)18.
- Doctor, S. R., J. D. Deffenbaugh, M. S. Good, E. R. Green, P. G. Heasler, F. A. Simonen, J. C. Spanner, and T. T. Taylor. 1988. Nondestructive Examination (NDE) Reliability for Inservice Inspection of Light Water Reactors, Semi-Annual Report, April-September 1987. NUREG/CR-4469, PNL-5711, Vol. 7, prepared by Pacific Northwest Laboratory for the U.S. Nuclear Regulatory Commission, Washington, D.C.
- Heasler, P. G., D. J. Bates, T. T. Taylor, and S. R. Doctor. 1986. Performance Demonstration Tests for Detection of Intergranular Stress Corrosion Cracking. NUREG/CR-4664, prepared by Pacific Northwest Laboratory for the U.S. Nuclear Regulatory Commission, Washington, D.C.
- Heasler, P. G., T. T. Taylor, J. C. Spanner, S. R. Doctor, and J. D. Deffenbaugh. 1990. Ultrasonic Inspection Reliability for Intergranular Stress Corrosion Cracks, A Round Robin Study of the Effects of Personnel, Procedures, Equipment, and Crack Characteristics. NUREG/CR-4098, prepared by Pacific Northwest Laboratory for the U.S. Nuclear Regulatory Commission, Washington, D.C.
- Hemsworth, B. 1986. "Inspection Trials - A Regulatory View." Presented at the PISC Symposium on Results sponsored by the British Institute of NDT, Daresbury, Cheshire, United Kingdom, November 25-26, 1986.
- McCullagh, P. and J. A. Nelder. 1983. Generalized Linear Models. Chapman & Hall, New York, New York.
- Murgatroyd, R. A., S. Crutzen, C. Vinche, and A. A. McFarlane. 1988. The Results of the Retrospective Analysis of the PISC-II Inspection Data. PISC-II Report #12. Ispra, Italy.
- Nichols, R. W., and S. Crutzen. 1988. Ultrasonic Inspection of Heavy Section Steel Components, The PISC-II Final Report. Elsevier Applied Science, New York, New York.
- Taylor, T. T., J. C. Spanner, P. G. Heasler, S. R. Doctor, and J. D. Deffenbaugh. July 1988. "An Evaluation of Human Reliability in Ultrasonic Inservice Inspection for Intergranular Stress Corrosion Cracks Through Round Robin Testing." Materials Evaluation.

Appendix A. Flaw and Grading Unit Plots

R9204050.45

R9204050.46

Flaws and Grading Units in Plate #1

Flaws and Grading Units in Plate #2

Appendix A

Flaws and Grading Units in Nozzle #3

Flaws and Grading Units in Nozzle #9

Appendix B. POD Curve Fits

The plots in this appendix are the POD versus flaw depth or flaw length. These regression plots are for advanced, manual-10% DAC, manual-20% DAC, and manual-50% DAC procedures. These plots include cladding and base material and the flaw types (volumetric, rough, and smooth).

POD vs. Flaw Depth: Base Material - Advanced Procedure - Rough Flaw

POD vs. Flaw Depth: Base Material - Advanced Procedure - Smooth Flaw

Appendix B

POD vs. Flaw Depth: Base Material - Advanced Procedure - Volumetric Flaw

POD vs. Flaw Depth: Clad Material - Advanced Procedure - Rough Flaw

POD vs. Flaw Depth: Clad Material - Advanced Procedure - Smooth Flaw

POD vs. Flaw Depth: Clad Material - Advanced Procedure - Volumetric Flaw

Appendix B

POD vs. Flaw Depth: Base Material - Manual-10% DAC Procedure - Rough Flaw

POD vs. Flaw Depth: Base Material - Manual-10% DAC Procedure - Smooth Flaw

POD vs. Flaw Depth: Base Material - Manual-10% DAC Procedure - Volumetric Flaw

POD vs. Flaw Depth: Clad Material - Manual-10% DAC Procedure - Rough Flaw

Appendix B

POD vs. Flaw Depth: Clad Material - Manual-10% DAC Procedure - Smooth Flaw

POD vs. Flaw Depth: Clad Material - Manual-10% DAC Procedure - Volumetric Flaw

POD vs. Flaw Depth: Base Material - Manual-20% DAC Procedure - Rough Flaw

POD vs. Flaw Depth: Base Material - Manual-20% DAC Procedure - Smooth Flaw

Appendix B

POD vs. Flaw Depth: Base Material - Manual-20% DAC Procedure - Volumetric Flaw

POD vs. Flaw Depth: Clad Material - Manual-20% DAC Procedure - Rough Flaw

POD vs. Flaw Depth: Clad Material - Manual-20% DAC Procedure - Smooth Flaw

POD vs. Flaw Depth: Clad Material - Manual-20% DAC Procedure - Volumetric Flaw

Appendix B

POD vs. Flaw Depth: Base Material - Manual-50% DAC Procedure - Rough Flaw

POD vs. Flaw Depth: Base Material - Manual-50% DAC Procedure - Smooth Flaw

POD vs. Flaw Depth: Base Material - Manual-50% DAC Procedure - Volumetric Flaw

POD vs. Flaw Depth: Clad Material - Manual-50% DAC Procedure - Rough Flaw

POD vs. Flaw Depth: Clad Material - Manual-50% DAC Procedure - Smooth Flaw

POD vs. Flaw Depth: Clad Material - Manual-50% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Base Material - Advanced Procedure - Rough Flaw

POD vs. Flaw Length: Base Material - Advanced Procedure - Smooth Flaw

POD vs. Flaw Length: Base Material - Advanced Procedure - Volumetric Flaw

POD vs. Flaw Length: Clad Material - Advanced Procedure - Rough Flaw

POD vs. Flaw Length: Clad Material - Advanced Procedure - Smooth Flaw

POD vs. Flaw Length: Clad Material - Advanced Procedure - Volumetric Flaw

Appendix B

POD vs. Flaw Length: Base Material - Manual-10% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Base Material - Manual-10% DAC Procedure - Smooth Flaw

POD vs. Flaw Length: Base Material - Manual-10% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Clad Material - Manual-10% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Clad Material - Manual-10% DAC Procedure - Smooth Flaw

POD vs. Flaw Length: Clad Material - Manual-10% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Base Material - Manual-20% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Base Material - Manual-20% DAC Procedure - Smooth Flaw

Appendix B

POD vs. Flaw Length: Base Material - Manual-20% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Clad Material - Manual-20% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Clad Material - Manual-20% DAC Procedure - Smooth Flaw

POD vs. Flaw Length: Clad Material - Manual-20% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Base Material - Manual-50% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Base Material - Manual-50% DAC Procedure - Smooth Flaw

POD vs. Flaw Length: Base Material - Manual-50% DAC Procedure - Volumetric Flaw

POD vs. Flaw Length: Clad Material - Manual-50% DAC Procedure - Rough Flaw

POD vs. Flaw Length: Clad Material - Manual-50% DAC Procedure - Smooth Flaw

POD vs. Flaw Length: Clad Material - Manual-50% DAC Procedure - Volumetric Flaw

Appendix C. Statistical Methods

Appendix C. Statistical Methods

The χ -squared statistics are generated by fitting a log-linear model to the detection data. The result is a table of χ -squared statistics that describe the effect of the variable (or variables) under consideration. The log-linear tables that we will use to evaluate these variables will contain two values, a χ -squared statistic and its associated degrees of freedom.

The χ -squared "goodness-of-fit" statistic measures whether or not a set of variables has a significant effect on POD. If the variables do not affect POD, the χ -squared statistic should be small and comparable in magnitude to its degrees of freedom. When the statistic exceeds its degrees of freedom by a significant amount, then the effect of the variables is not likely being confused with the effects of random variation. A χ -squared table can be used to determine the magnitude at which the statistic becomes significant.

These goodness-of-fit statistics are generated by fitting a model with the mathematical form

$$E(D_{jk}) = \exp (\mu + \alpha_j + \beta_k + \gamma_{jk})$$

where j = levels of the variable(s) that we want to test

k = levels of the "control" variables

D_{jk} = detection variable. (It is 1 if detection occurred under conditions jk and 0 if detection did not occur.)

The assertion that variable j has no effect on detection is equivalent to saying that all terms involving the variable j can be eliminated from the formula above. The χ -squared "goodness-of-fit" statistic is designed to test whether these terms are "significant." The statistic is called a "goodness-of-fit" statistic because it is calculated by fitting the above equation to the data and comparing this fit obtained by fitting the reduced model

$$E(D_{jk}) = \exp (\mu \beta_k)$$

Appendix D. Sizing Curve Fits

Measured Depth Vs. True Depth: Base Material - 20%-DAC Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Base Material - 20%-DAC Amplitude Drop - Rough Flaw

Appendix D

Measured Depth Vs. True Depth: Base Material - 20%-DAC Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Base Material - 20%-DAC Amplitude Drop - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - 20%-DAC Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Base Material - 20%-DAC Amplitude Drop - Volumetric Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - 20%-DAC Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Clad Material - 20%-DAC Amplitude Drop - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - 20%-DAC Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Clad Material - 20%-DAC Amplitude Drop - Smooth Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - 20%-DAC Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - 20%-DAC Amplitude Drop - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - 50%-DAC Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Base Material - 50%-DAC Amplitude Drop - Rough Flaw

Appendix D

Measured Depth Vs. True Depth: Base Material - 50%-DAC Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Base Material - 50%-DAC Amplitude Drop - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - 50%-DAC Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Base Material - 50%-DAC Amplitude Drop - Volumetric Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - 50%-DAC Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Clad Material - 50%-DAC Amplitude Drop - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - 50%-DAC Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Clad Material - 50%-DAC Amplitude Drop - Smooth Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - 50%-DAC Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - 50%-DAC Amplitude Drop - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Base Material - Amplitude Drop - Rough Flaw

Measured Depth Vs. True Depth: Base Material - Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Base Material - Amplitude Drop - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Base Material - Amplitude Drop - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Amplitude Drop - Rough Flaw

Measured Length Vs. True Length: Clad Material - Amplitude Drop - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - Amplitude Drop - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Amplitude Drop - Smooth Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - Amplitude Drop - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - Amplitude Drop - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Combination of Techniques - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Combination of Techniques - Smooth Flaw

Measured Depth Vs. True Depth: Clad Material - Combination of Techniques - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - Combination of Techniques - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - Tip Diffraction - Rough Flaw

Measured Length Vs. True Length: Base Material - Tip Diffraction - Rough Flaw

Appendix D

Measured Depth Vs. True Depth: Base Material - Tip Diffraction - Smooth Flaw

Measured Length Vs. True Length: Base Material - Tip Diffraction - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - Tip Diffraction - Volumetric Flaw

Measured Length Vs. True Length: Base Material - Tip Diffraction - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Tip Diffraction - Rough Flaw

Measured Length Vs. True Length: Clad Material - Tip Diffraction - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - Tip Diffraction - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Tip Diffraction - Smooth Flaw

Measured Depth Vs. True Depth: Clad Material - Tip Diffraction - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - Tip Diffraction - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - Focused Technique - Rough Flaw

Measured Length Vs. True Length: Base Material - Focused Technique - Rough Flaw

Measured Depth Vs. True Depth: Base Material - Focused Technique - Smooth Flaw

Measured Length Vs. True Length: Base Material - Focused Technique - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - Focused Technique - Volumetric Flaw

Measured Length Vs. True Length: Base Material - Focused Technique - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Focused Technique - Rough Flaw

Measured Length Vs. True Length: Clad Material - Focused Technique - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - Focused Technique - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Focused Technique - Smooth Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - Focused Technique - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - Focused Technique - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - Computer Imaging Technique - Rough Flaw

Measured Length Vs. True Length: Base Material - Computer Imaging Technique - Rough Flaw

Measured Depth Vs. True Depth: Base Material - Computer Imaging Technique - Smooth Flaw

Measured Depth Vs. True Depth: Base Material - Computer Imaging Technique - Volumetric Flaw

Measured Length Vs. True Length: Base Material - Computer Imaging Technique - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Computer Imaging Technique - Rough Flaw

Measured Length Vs. True Length: Clad Material - Computer Imaging Technique - Rough Flaw

Measured Depth Vs. True Depth: Clad Material - Computer Imaging Technique - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Computer Imaging Technique - Smooth Flaw

Measured Depth Vs. True Depth: Clad Material - Computer Imaging Technique - Volumetric Flaw

Measured Length Vs. True Length: Clad Material - Computer Imaging Technique - Volumetric Flaw

Measured Depth Vs. True Depth: Base Material - Simple Imaging Technique - Rough Flaw

Measured Length Vs. True Length: Base Material - Simple Imaging Technique - Rough Flaw

Appendix D

Measured Depth Vs. True Depth: Base Material - Simple Imaging Technique - Volumetric Flaw

Measured Length Vs. True Length: Base Material - Simple Imaging Technique - Volumetric Flaw

Measured Depth Vs. True Depth: Clad Material - Simple Imaging Technique - Rough Flaw

Measured Length Vs. True Length: Clad Material - Simple Imaging Technique - Rough Flaw

Appendix D

Measured Depth Vs. True Depth: Clad Material - Simple Imaging Technique - Smooth Flaw

Measured Length Vs. True Length: Clad Material - Simple Imaging Technique - Smooth Flaw

**Appendix E. ASME Section XI Code Performance Demonstration
Requirements for Clad/Base-Metal Region
and Pressure Vessel Shell Welds**

Appendix E. ASME Section XI Code Performance Demonstration Requirements for Clad/Base-Metal Region and Pressure Vessel Shell Welds

APPENDIX VIII

Performance Demonstration Requirements for the Clad/Base Metal Region of Reactor Vessel

1.0 SPECIMEN REQUIREMENTS

Specimens to be used in the qualification test shall meet the requirements listed herein unless a set of test specimens is designed to accommodate specific limitations stated in the scope of the examination procedure. The same specimens may be used to demonstrate both detection and sizing qualifications.

1.1 Detection Specimens

Detection specimens, which may be full-scale mock-ups, shall conform to the following requirements:

- (a) Specimens shall have sufficient area to minimize spurious reflections, which may interfere with the interpretation process. Specimen length and width shall be at least 12 inches. There shall be at least ten square feet of clad surface in the specimen set.
- (b) (1) When the examination procedure requires the examination to be performed from the vessel I.D. (clad surface), the specimen minimum thickness shall be three inches or the maximum thickness of the vessel (whichever is less).
- (2) When the examination procedure requires the examination to be performed from the vessel O.D. surface, the specimen shall be at least the maximum thickness of the vessel components to be examined.
- (c) The performance demonstration shall be on the same type cladding as that to be examined, with the following exceptions:
 - (1) Demonstration on shielded metal arc weld (SMAW) single-wire cladding is transferable to multiple-wire or strip-clad processes.
 - (2) Demonstration of multiple-wire or strip-clad are considered equivalent but is not transferable to SMAW-type clad.
- (d) The surface condition of the test specimens shall be representative of the general condition of the vessel scanning surface.
- (e) The detection test matrix shall include flaws with the following description:
 - (1) Flaw Type - At least 50% of the flaws shall be cracks. The balance of flaws may be cracks, fabrication defects such as slag inclusions, or machined notches. Machined notches shall meet the following requirements:
 - (a) Notches shall be perpendicular to the surface within ± 2 degrees.
 - (b) Notches shall have a maximum width of 0.010 inch and shall be plugged to their full depth with an insert of the parent material prior to cladding.
 - (2) All flaws shall emanate from the clad base metal interface and shall propagate predominantly into the base metal. A minimum of 40% of the flaws shall be oriented parallel to the clad direction (within ± 10 degrees) and a minimum of 40% shall be oriented perpendicular to the clad direction (within ± 10 degrees).
 - (3) The flaw sizes shall be uniformly distributed in through-wall depths (inches) among the ranges: a) 0.075-0.200, b) 0.201-0.350, c) 0.351-0.550, and d) 0.551-0.750. No flaw shall have an aspect ratio (depth/length) less than 0.1.
- (f) The number of flaws in the detection test matrix shall be selected from Table VIII-S4-1. The rules of IWA-3500 shall be used for determining whether closely spaced flaws should be treated as separate flaws.

1.2 Sizing Specimens

- (a) The sizing test matrix shall contain a minimum of ten flaws, at least half of which shall be cracks.
- (b) Sizing specimens shall conform with the requirements of 1.1(b), (c), (d), and (e).

2.0 CONDUCT OF PERFORMANCE DEMONSTRATIONS

2.1 Detection Test

- (a) Specimen identification and flaw locations shall be obscured so as to maintain a "blind test." All examinations shall be completed prior to grading the results and presenting the results to the candidate. Divulgence of particular specimen results or candidate viewing of unmasked specimens is prohibited.
- (b) When a flaw is reported within 1/2 inch of its true location (x and y) it shall be considered detected. All other reported flaws shall be considered false calls.

2.2 Length and Depth Sizing Test

- (a) For the length sizing test, the inside surface and specimen identification shall be obscured. The regions of each specimen containing a flaw to be sized shall be identified to the candidate. The candidate shall determine the length of the flaw in each region.
- (b) For the depth sizing test, 80% of the flaws shall be sized at a specific location on the surface of the specimen identified to the candidate. For the remaining flaws, the regions of each specimen containing a flaw to be sized shall be identified to the candidate. The candidate shall determine the maximum depth of the flaw in each region.

3.0 ACCEPTANCE CRITERIA

3.1 Detection Acceptance Criteria

Examination procedures, equipment, and personnel shall be considered qualified for detection if the results of the performance demonstration satisfy the acceptance criteria of Table VIII-S4-1. Additionally, the number of false calls shall not exceed $A/10$ rounded to the next whole number where A is the total scan area of specimens in the test measured in square feet.

3.2 Sizing Acceptance Criteria

Examination procedures, equipment, and personnel shall be considered qualified for sizing if the results of the performance demonstration satisfy the following criteria:

- (a) No flaw is undersized for depth by more than 0.2 inches.
- (b) Flaw lengths estimated by ultrasonics shall be the true length $-1/4$ inch, $+1$ inch.
- (c) Performance demonstration results reported by the candidate, when plotted on a two-dimensional plot (Figure VIII-S2-1) with the depth estimated by ultrasonics plotted along the ordinate and the true depth plotted along the abscissa, satisfy the following statistical parameters:
 - (1) Slope of the linear regression line is not less than 0.7.
 - (2) The mean deviation of flaw depth is less than 0.25 inch.
 - (3) Correlation coefficient is not less than 0.70.

Table VIII-S4-1

**Performance Demonstration Detection
Test Acceptance Criteria**

**DETECTION TEST
ACCEPTANCE CRITERIA**

No. of Flaws	Minimum Detection Criteria
7	7
8	8
9	9
10	10
11	11
12	11
13	12
14	13
15	14
16	14
17	15
18	16
19	17
20	18

APPENDIX VIII

**Performance Demonstration Requirements for
Reactor Pressure Vessel Shell Welds Other
Than Clad/Base Metal Region**

1.0 SPECIMEN REQUIREMENTS

Specimens to be used in the qualification test shall meet the requirements listed herein unless a set of test specimens is designed to accommodate specific limitations stated in the scope of the examination procedure. The same specimens may be used to demonstrate both detection and sizing qualification.

1.1 Detection Specimens

Detection specimens, which may be full-scale mock-ups, shall conform to the following requirements:

- (a) Specimens shall have sufficient area to minimize spurious reflections, which may interfere with the interpretation process. Specimen length and width shall be at least 12 inches. There shall be at least ten square feet of scan surface in the specimen set.
- (b) The specimen set shall be at least the maximum thickness of the vessel components to be examined.
- (c) When the examination procedure requires the examination to be performed from the vessel I.D. (clad surface), the performance demonstration shall be on the same type cladding I.D. surface. If the vessel I.D. surface is clad, the performance demonstration shall be on the same type cladding as that to be examined, with the following exceptions:
 - (1) Demonstration on shielded metal arc weld (SMAW) single-wire cladding is transferable to multiple-wire or strip-clad processes.
 - (2) Demonstration of multiple-wire or strip-clad are considered equivalent but is not transferable to SMAW-type clad.

- (d) The surface condition of the test specimens shall be representative of the general condition of the vessel scanning surface.
- (e) The detection test matrix shall include flaws with the following description:
 - (1) Flaw Type - At least 50% of the flaws shall be cracks. The balance of flaws may be cracks, fabrication defects such as slag inclusions, or machined notches. Machined notches shall meet the following requirements:
 - (a) Notches shall be perpendicular to the surface within ± 2 degrees.
 - (b) Notches shall have a maximum width of 0.010 inch and shall be plugged to their full depth with an insert of the parent material.
 - (2) No more than 10% of the flaws shall emanate from the outer ~~(unexposed)~~ surface. A weld direction shall be established whether or not the specimen contains a weld.

A minimum of 40% of the flaws shall be oriented parallel to the weld direction (within ± 10 degrees) and a minimum of 40% shall be oriented perpendicular to the weld direction (within ± 10 degrees).
 - (3) The location of flaws shall be distributed as shown in Table VIII-S6-1. Flaw sizes shall be uniformly distributed among applicable flaw size ranges shown in Table VIII-S6-1.
 - (4) The number of flaws in the detection test matrix shall be selected from Table VIII-S6-2. The rules of IWA-3500 shall be used for determining whether closely spaced flaws should be treated as separate flaws.

1.2 Sizing Specimens

- (a) The sizing test matrix shall contain a minimum of ten flaws, at least half of which shall be cracks.

- (b) Sizing specimens shall conform with the requirements of 1.1(b), (c), (d), and (e).

2.0 CONDUCT OF PERFORMANCE DEMONSTRATIONS

2.1 Detection Test

- (a) Specimen identification and flaw locations shall be obscured so as to maintain a "blind test." All examinations shall be completed prior to grading the results and presenting the results to the candidate. Divulgence of particular specimen results or candidate viewing of unmasked specimens is prohibited.
- (b) When a flaw is reported within 1/2-inch of its true location (x and y), it shall be considered detected. All other reported flaws shall be considered false calls.

2.2 Length and Depth Sizing Test

- (a) For the length sizing test, the inside surface and specimen identification shall be obscured. The regions of each specimen containing a flaw to be sized shall be identified to the candidate. The candidate shall determine the length of the flaw in each region.
- (b) For the depth sizing test, 80% of the flaws shall be sized at a specific location on the surface of the specimen identified to the candidate. For the remaining flaws, the regions of each specimen containing a flaw to be sized shall be identified to the candidate. The candidate shall determine the maximum depth of the flaw in each region.

3.0 ACCEPTANCE CRITERIA

3.1 Detection Acceptance Criteria

Examination procedures, equipment, and personnel shall be considered qualified for detection if the results of the performance demonstration satisfy the acceptance criteria of Table VIII-S6-2. Additionally, the number of false calls shall not exceed $A/10$ rounded to the next whole number where A is the total

Table VIII-S6-1
Detection Test Flaw Locations

Through-Wall Dimension, cm	A	B	C	D
	0.075-0.200	0.201-0.350	0.351-0.550	0.551-0.750
Inside Surface*	X	X		
Outside Surface	X	X		
10-30% T			X	X
31-60% T			X	X
61-90% T			X	X

*Does not apply to clad vessels (see Supplement 4).

scan area of specimens in the test measured in square feet.

3.2 Sizing Acceptance Criteria

Examination procedures, equipment, and personnel shall be considered qualified for sizing if the results of the performance demonstration satisfy the following criteria:

(a) No flaw is undersized for depth by more than 0.2 inches. (b) Flaw lengths estimated by ultrasonics shall be the true flaw length -1/4 inch, +1 inch.	(c) Performance demonstration results reported by the candidate, when plotted on a two-dimensional plot (Figure VIII-S2-1) with the depth estimated by ultrasonics plotted along the ordinate and the true depth plotted along the abscissa, satisfy the following statistical parameters: (1) Slope of the linear regression line is not less than 0.70. (2) The mean deviation of flaw depth is less than 0.25 inch. (3) Correlation coefficient is not less than 0.70.

Table VIII-S6-2

**Performance Demonstration Detection
Test Acceptance Criteria**

**DETECTION TEST
ACCEPTANCE CRITERIA**

No. of Flaws	Minimum Detection Criteria
7	7
8	8
9	9
10	10
11	11
12	11
13	12
14	13
15	14
16	14
17	15
18	16
19	17
20	18

DISTRIBUTION

No. of
Copies

No. of
Copies

OFFSITE

FOREIGN

2 J. Muscara
NRC/RES
Mail Stop NS 217C

C. Z. Serpan
NRC/RES
Mail Stop NS 217C

G. Johnson
NRC/RES
Mail Stop 7D-4

M. R. Hum
NRC/NRR
Mail Stop 7D-4

J. T. Wiggins
NRC/NRR
Mail Stop 7D-4

J. P. Durr
NRC/Region I

M. C. Modes
NRC/Region I

A. R. Herdt
NRC/Region II

J. J. Blake
NRC/Region II

K. Ward
NRC/Region III

J. R. Tomlinson
NDT Application Centre
Nuclear Electric plc
Timpson Road
Wythenshawe
Manchester M23 9LL
United Kingdom

45 S. Crutzen
Joint Research Centre, Ispra
I-21020 Varese
Italy

ONSITE

50 Pacific Northwest Laboratory

E. S. Andersen
R. E. Bowey
S. H. Bush
A. A. Diaz
S. R. Doctor (30)
J. R. Friley
B. F. Gore
M. S. Greenwood
R. V. Harris
P. G. Heasler
R. J. Kurtz
F. A. Simonen
J. C. Spanner
T. T. Taylor
T. V. Vo
Publishing Coordination
Technical Report Files (5)

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-5410
PNL-8577

2. TITLE AND SUBTITLE

Statistically Based Reevaluation of PISC-II
Round Robin Test Data

3. DATE REPORT PUBLISHED

MONTH | YEAR
May | 1993

4. FIN OR GRANT NUMBER

B2289

5. AUTHOR(S)

P.G. Heasler, T.T. Taylor, S.R. Doctor

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 Laboratory
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

10. SUPPLEMENTARY NOTES

11. ABSTRACT (200 words or less)

This report presents a re-analysis of an international PISC-II round-robin inspection results using formal statistical techniques to account for experimental error. The analysis examines:

- U.S. team performance vs. other participants performance
- flaw sizing performance and errors associated with flaw sizing
- factors influencing flaw detection probability
- performance of all participants with respect to recently developed ASME Section XI flaw detection performance demonstration requirements

and develops conclusions concerning ultrasonic inspection capability.

12. KEY WORDS, DESCRIPTORS (List words or phrases that will assist researchers in locating the report.)

nondestructive evaluation, nondestructive testing, thick section steel, fracture mechanics, PISC II, ultrasonic testing, ASME Code, performance demonstration, reactor pressure vessel, nozzles, radiography

13. AVAILABILITY STATEMENT

Unlimited

14. SECURITY CLASSIFICATION

(This Page)

Unclassified

(This Report)

Unclassified

15. NUMBER OF PAGES

16. PRICE

Federal Recycling Program