

## METHODOLOGY 4

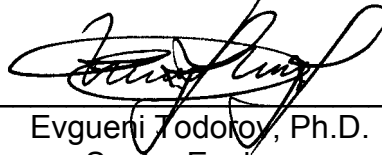
### Methodology for Data Analysis

Issued by:  
Edison Welding Institute

Revision: 5

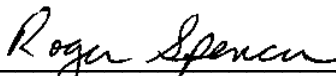
Revision Date: February 18, 2011

Prepared by:



Evgueni Todokov, Ph.D.  
Senior Engineer  
Nondestructive Evaluation

Approved by:



Roger Spencer  
Senior Engineer  
Nondestructive Evaluation

February 18, 2011  
Date Approved

## Methodology for Data Analysis

### 1.0 Scope

1.1 This methodology describes major steps and requirements in the data analysis for estimation of sizing and detection capabilities of automated ultrasonic testing (AUT) systems for nondestructive evaluation (NDE) inspection of girth welds.<sup>(1)</sup>

1.2 The assessment of sizing capabilities can be done separately and independently of the assessment of detection capabilities.

1.3 A destructive testing (DT)<sup>(2)</sup> is used as reference technique to establish the AUT detection and sizing capabilities. The results from other measurement techniques (NDE including) with better detection capabilities and flaw sizing errors significantly smaller than the AUT system errors can be used as reference.

### 2.0 Flaw Sizing Definitions

2.1 Where a flaw is detected, the AUT system provides estimate of the flaw dimensions such as height, length, depth, and position (start and stop) along the circumference. Standard guidelines are followed to describe and express the uncertainty of the measurements.<sup>(3)</sup>

2.2 The measurand is the particular quantity subject to measurement. For AUT of girth welds, the measurand is the flaw height, length, depth, start, and stop position.

2.3 Any single AUT measurement provides an estimate  $\hat{y}_i$  of the measurand consisting of its “true” value  $a_i$  and a measurement error  $\varepsilon_i$  in accordance with Eq. (1):

$$(1) \quad \hat{y}_i = a_i + \varepsilon_i$$

2.4 The error is usually assumed to be normally distributed with a mean and standard deviation. For analysis purposes, the error  $\varepsilon_i$  consists of a systematic ( $Sys \varepsilon_i$ ) and a random ( $Ran \varepsilon_i$ ) component<sup>(4)</sup> shown in Eq. (2):

$$(2) \quad \varepsilon_i = Sys \varepsilon_i + Ran \varepsilon_i$$

2.5 The “true” value  $a_i$  of the measurand and the error  $\varepsilon_i$  are never known and can only be estimated. The “true” value estimate is provided by a metallographic DT test or other more accurate reference method where available. An estimate of a single measurement error is the difference between the AUT estimate and the reference (“true”) measurand value Eq. (3):

$$(3) \quad \varepsilon_i = \hat{y}_i - a_i$$

2.6 An estimate of the error systematic component ( $Est (Sys \varepsilon_i)$ ) in Eq. (2) is provided by averaging the individual errors of a large number of  $n$  measurements shown in Eq. (4) below:

$$(4) \quad \bar{\varepsilon} = Est(Sys \ \varepsilon_i) = \frac{\sum_{i=1}^n \varepsilon_i}{n}$$

2.7 An estimate of the random error spread or dispersion is provided by the error standard deviation  $s(\varepsilon)$  and variance  $V(\varepsilon)$  shown in Eq. (5). The standard deviation is also referred to as Standard Uncertainty.<sup>(3)</sup>

$$(5) \quad s(\varepsilon) = \sqrt{V(\varepsilon)} = \sqrt{\frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2}{n-1}}$$

2.8 For height and length AUT measurements, the 95% safety limit against undersizing ( $95\%LUS$ ) is shown in Eq. (6) where  $k$  is a coverage factor.<sup>(3)</sup> If the error is normally distributed, the coverage factor becomes the standardized normal deviate<sup>(5)</sup> with value of 1.645 for large number ( $n > 120$ ) of measurements. For normally distributed data with  $n < 120$ , a parameter  $t$  from Student's distribution with 95% (5% one tail) probability should be used as coverage factor. The value of the parameter  $t$  is determined from statistical tables or dedicated statistical software (e.g., Minitab®, MS Excel®).

$$(6) \quad 95\%LUS = ks(\varepsilon) - \bar{\varepsilon}$$

2.9 The term  $ks(\varepsilon)$  is referred to as Expanded Uncertainty<sup>(3)</sup> defining an interval expected to encompass a fraction of the distribution of values attributed to the measurand at a specified confidence level.

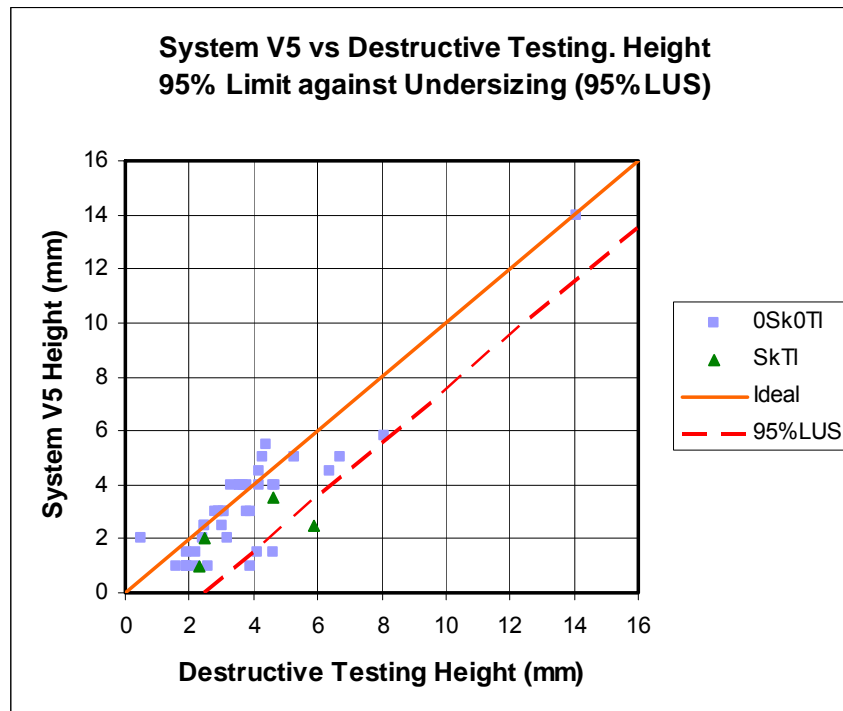
2.10 The ( $95\%LUS$ ) is also referred to as 5% error fractile or undersizing error tolerance that gives equal or less than 5% probability of undersizing.<sup>(6)</sup> The estimation of  $95\%LUS$  requires knowledge of the distribution law (probability density function) so that a value of the coverage factor can be obtained for the desired confidence.

2.11 For AUT depth measurements, the standard form<sup>(3)</sup> of expressing the results from each depth measurements is used accounting for the systematic error and the expanded uncertainty interval as shown in Eq. (7):

$$(7) \quad d_i = \hat{d}_i - \bar{\varepsilon} \pm ks(\varepsilon)$$

2.12 The “true” depth value  $d_i$  is then expected to be in an interval as shown in Eq. (8):

$$(8) \quad \hat{d}_i - \bar{\varepsilon} - ks(\varepsilon) \leq d_i \leq \hat{d}_i - \bar{\varepsilon} + ks(\varepsilon)$$



**Figure 1. Typical Sizing Plot with Ideal and 95%LUS Lines.**

2.13 A typical sizing plot ( $95\%LUS > 0$ ) comparing AUT and actual flaw measurements is shown in Figure 1. The closer the AUT measurement and 95%LUS are to the “ideal” line the better the system sizing performance is.

2.14 The error of a single height measurement  $\varepsilon_i$  is estimated by comparing the maximum of the “true” height  $a_i$  to the maximum of the AUT estimate  $\hat{y}_i$  of the height. The depth error estimates are obtained by comparing the “true” to the AUT depth for the flaw location where the maximum flaw height was measured.

### 3.0 Procedure for Sizing Data Analysis

3.1 The sizing error components Eqs. (2)-(5) need to be estimated to compare the weld specimens, systems, processes, and assess the AUT system sizing performance and capabilities.

3.2 Each sizing error data sample should be processed in accordance with the following procedure:

- Obtain average
- Obtain standard deviation (variance) or uncertainty
- Plot histogram and compare to normal distribution
- Perform normality test (Anderson-Darling)
- Build and analyze box plots
- Obtain and analyze other statistics - kurtosis, skewness, range, confidence intervals (CI) etc.

- Perform equal variance test
- Perform analysis of variance (ANOVA) or nonparametric tests as applicable to check whether a statistically significant difference exists between distributions
- Identify outliers
- Perform parametric Test 1 – “2-Sample t” for two or “F-Tukey's” for more than two distributions that are normally distributed and have equal variance
- Perform nonparametric Test 2 – “Mann-Whitney” for two or “Kruskal-Wallis” for more than two distributions that are not normally distributed and or do not have equal variance.

3.3 A standard statistical significance for all tests (P-Value) of 0.05 (or 5%) is recommended. However, the P-Value should not be used as an ultimate criteria to accept or reject a statistical hypothesis of data normality and distribution similarity. Additional tests, analysis and past experience might be considered where possible.

3.4 Outliers affect the average, standard uncertainty, and distribution estimates and might be removed if the effect is significant and the sample size is not reduced by the outliers removal to a level of being unrepresentative. The removed outliers should be documented along with the reason for removal.

3.5 Some flaw types and or weld-specimen group data samples might be removed from the joint data samples if statistically significant differences in error distributions are observed. The joint sample size, however, should not be significantly affected by the group sample removal. The removed data group should be documented along with the reason for removal.

#### **4.0 Probability of Detection Definitions and Methods**

4.1 Probability of flaw detection describes the AUT system capability to detect flaws. There are four outcomes of any inspection: true positive, true negative, false negative or miss and false positive or false alarm.

4.2 Three parameters are usually obtained as a result of dedicated POD studies -  $a_{50}$ ,  $a_{90}$ , and  $a_{90/95}$ . The interpretation or definition of each of these three parameters is as follows:

- $a_{50}$  – flaw size with 50% POD. This means that 50% of the flaws with this size and larger will be detected.
- $a_{90}$  – flaw size with 90% POD. This means that 90% of the flaws with this size and larger will be detected.
- $a_{90/95}$  – flaw size with 90% POD and 95% confidence. This is the most quoted parameter in the literature. It means that 90% of the flaws with this size and larger will be detected and this is true in 95% of the inspections under similar conditions (equipment, examiners, environment, etc.).

4.3 Option 1 – Binomial test<sup>(4,7)</sup> for a single flaw size.

4.3.1 The POD(a) expressed as  $a_{90/95}$  can be estimated for a single flaw size applying the so called “29-out-of-29” rule. It means that 29 out of 29 flaws at a given size must be detected to demonstrate a  $a_{90/95}$  at this size.

4.3.2 It is difficult to fabricate flaws with identical size and the actual flaws will be expected to cover an interval of sizes. The  $a_{90/95}$  will then be the largest flaw size in the range for which the 29-out-of-29 rule applies. This approach might be applicable for NDT systems quantification where the  $a_{90/95}$  is known and it is in the range of the AUT system capabilities. It can also be used for fast comparison and selection of AUT systems for a particular application.

4.3.3 The number of detected and missed flaws with typical POD and confidence values is shown in Table 1. A more detailed table is available in the literature.<sup>(4)</sup>

**Table 1. Minimum Number of Weld Sectors with Flaws for a Given POD, Confidence and Number of Misses**

Confidence	Number of Misses	Number of Weld Sectors with Flaws		
		POD 80%	POD 90%	POD 95%
50%	0	3	7	14
	1	7	17	32
	2	11	27	51
	3	15	37	70
	4	19	47	89
	5	23	57	108
	10	43	107	203
	20	83	207	394
90%	0	11	22	45
	1	17	38	77
	2	23	52	105
	3	29	65	132
	4	34	78	158
	5	39	91	184
	10	64	152	306
	20	112	267	538
95%	0	13	29	59
	1	21	46	93
	2	27	61	124
	3	33	76	153
	4	39	89	181
	5	45	103	208
	10	72	167	336
	20	121	286	577

4.3.4 If the AUT system misses flaws from the set, the number of flaws required to achieve  $a_{90/95}$  rapidly increases. For example, the minimum required number of 29 detected flaws increases to 45 or 59 flaws if 1 or 2 flaws, respectively, are missed (Table 1).

4.4 Option 2 –  $POD(a)$  curve [4, 8] for range of flaw sizes.

4.4.1 A binary regression is the recommended statistical method for obtaining of POD estimates and capabilities of AUT systems for a range of flaw sizes. There are two major approaches when building  $POD(a)$  curves – “ $\hat{a}$  vs  $a$ ” and “*hit/miss*”:

- When  $\hat{a}$  vs  $a$  (a hat versus a) approach is implemented,  $a$  is the actual flaw size and  $\hat{a}$  is the instrument response (e.g., millivolts, screen divisions, percent of screen height and others) to the flaw with size  $a$ .
- The *hit/miss* approach requires that only two conditions of instrument response are considered: *hit* (pass) – a flaw with size  $a$  was detected (instrument response coded as “1”) and *miss* (fail) – a flaw with size  $a$  was missed (instrument response coded as “0”).

4.4.2 Detailed explanation of assumptions, necessary conditions to use either  $\hat{a}$  vs  $a$  or *hit/miss* approach and data interpretation is outside of the scope of this methodology and can be found elsewhere.<sup>(8)</sup>

4.4.3 The most common function to approximate or model the behavior of  $POD(a)$  is the log-odds function<sup>(8)</sup> with probability transformation function shown in Eqs. (9) and (10):

$$(9) \quad POD(a) = \frac{e^Y}{1 + e^Y} \text{ logistic, log-odds or logit link}$$

$$(10) \quad Y = \log\left(\frac{p}{1-p}\right) \text{ logistic probability transformation}$$

4.4.4 The flaw size  $a$  is either transformed (e.g., log) Eq. (11) or not Eq. (12):

$$(11) \quad X = \log(a)$$

$$(12) \quad X = a$$

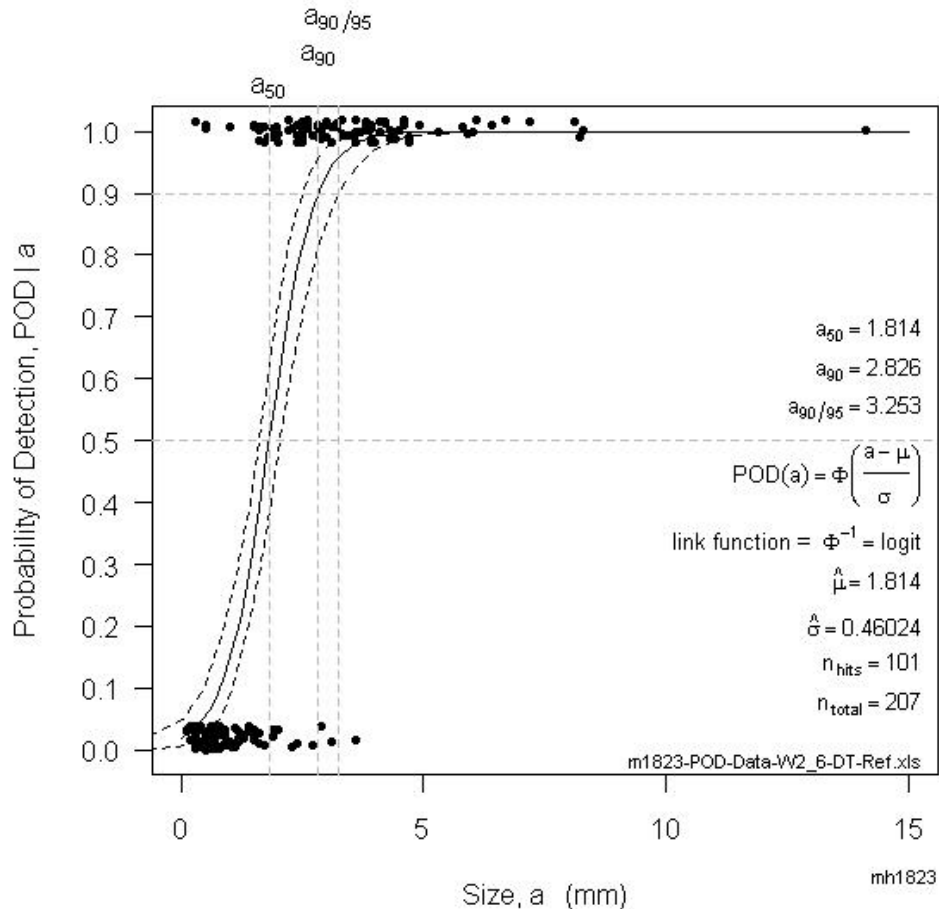
4.4.5 Other functions for approximation of  $POD(a)$  are also possible to use.<sup>(4,8)</sup> The minimum deviance criteria<sup>(8)</sup> is usually (not always) used to select the best link function and whether a flaw size transformation is needed or not.

4.4.6 A  $POD(a)$  plot for a typical AUT system with  $a_{50}$ ,  $a_{90}$ , and  $a_{90/95}$ , logit link function without flaw size transformation and other parameters is shown in Figure 2.

4.4.7 The quantification of various AUT systems and/or AUT examiners in terms of  $POD(a)$  should be conducted with the same or similar sets of specimens covering the same range of flaw sizes and build according to the requirements of this and other relevant methodologies. The detection results of several systems or examiners with similar performance demonstrated on the same flaw specimen set may be combined to build an average  $POD(a)$  for the inspection technique or team.

4.4.8 A quantification performed by different AUT examiners or AUT systems on the same flaw specimen set should not be interpreted as increase of the effective number of flaws in the set. For example, a quantification study performed by one examiner with a set of 60 flaws covering a

range of flaw sizes of interest is not equivalent to a quantification study performed by five examiners on another set with 12 flaws covering the same range of flaw sizes. The confidence in estimating  $POD(a)$  decreases when the number of flaws is small.<sup>(8)</sup>



**Figure 2. Typical AUT System  $POD(a)$  Plot with Confidence Intervals**

4.4.9 Although combining of data from AUT examiners or systems using different instrumentation or techniques [pulse-echo and time-of-flight diffraction (TOFD)] or weld specimens with different weld bevel preparation may improve the overall or average performance, it should be treated with caution. The performance of each AUT examiner or system should first be analyzed separately to identify whether combining of data is possible or not.

4.4.10 Different tools for estimation of  $POD(a)$  may produce different results because the estimates depends not only on the approximation or model function but on the way the flaw sizes are grouped into intervals, the way the confidence intervals for  $POD(a)$  are calculated and others. A consideration and selection of various approaches and related tools should be conducted before the quantification to ensure the  $POD(a)$  estimates are representative and conservative.



4.4.11 All parties related to the quantification process – the contracting party, administrator and AUT vendor should agree on the way the  $POD(a)$  estimates will be conducted and the results interpreted.

## 5.0 False Positive Frequency and Probability of False Positives

5.1 The false positive frequency (FPF) is estimated as shown in Eq. (13):

$$(13) \quad \text{False Positive Frequency} = \frac{D_{NF}}{N} \times 100\%,$$

where  $D_{NF}$  – number of weld sectors where false flaw indications are present,  $N$  – total number of inspected weld sectors.

5.2 If the weld sectors are not explicitly defined, the weld length can be divided into sectors with length at least equal to the ultrasonic beam width at the weld. For example, if the ultrasonic beam width parallel to the line of scanning is 25 mm at -6 dB level and the weld length is 2000 mm, the total number of weld sectors used to calculate the FPF is  $2000/50 = 40$ .

5.3 Another approach to estimate the false positive indications is to provide the number of false indications per unit of scanned length for the entire scanned weld length.<sup>(4,8)</sup> For example, if 3 false indications were detected per 12 m of weld length, the false positives will be 0.25 per 1 m. The economic and scheduling impact of false positives for a given AUT system or examiner during field inspections can easily be estimated by multiplying the false positives per unit length by the entire inspected weld length.

5.4 The probability of false positives (PFP) can also be estimated with specialized computational tools.<sup>(8)</sup> To estimate the PFP, the number of false positive indications and the total number (where available or as defined in 5.2 above) of all scanned weld sectors are used.

## 6.0 Reporting and Documentation

6.1 The results of the test evaluation and interpretation are documented in a report section that should contain as a minimum the following information.

6.1.1 Sizing curves (e.g., Figure 1) for each AUT examiner, AUT system and AUT vendor as applicable with 95%LUS line. Major statistical parameters characterizing the sizing accuracy such as 95%LUS, systematic or average error and standard uncertainty.

6.1.2 For POD Option 1, the POD achieved with POD confidence level for each AUT examiner, AUT system and AUT vendor as applicable.

6.1.3 For POD Option 2, the  $POD(a)$  function curves, with lower 95% confidence bound at least, and statistical model parameters (e.g., Figure 2) for each AUT examiner, AUT system and AUT vendor as applicable. Major POD parameters such as  $a_{50}$ ,  $a_{90}$ , and  $a_{90/95}$ .

6.1.4 An estimate of the FPF, false positives per unit length and or PFP as applicable to compare performance of different examiners, systems, and vendors.

6.1.5 Although not required, it is recommended that the sizing and detection results be further reported for different depths within the weld; for example: cap, fill, hot pass, and root.

6.2 All the data from the previous stages of the quantification process (flaw fabrication, specimen fingerprinting, AUT of girth welds, and destructive testing) shall be summarized and presented with the AUT errors and basic statistics in a summary table shown in Appendix 1.

6.3 The evaluation report will be part of the report submitted at the end of the open and blind trials and destructive testing (if required).

## **7.0 References**

- (1) Guidance for Quantification of Automated Ultrasonic Testing Systems for Examination of Pipeline Girth Welds, Edison Welding Institute (EWI).
- (2) Methodology for Practical Trials and Destructive Validation, Edison Welding Institute (EWI).
- (3) ANSI/NCSL Z540-2-1997 (R2007) "U.S. Guide to the Expression of Uncertainty in Measurement", American National Standard for Calibration, NCSL International.
- (4) Nordtest Report NT TECHN REPORT 394, Approved 1998-04
- (5) Box, G. E. P., J. S. Hunter, and W. G. Hunter, "Statistics for Experimenters: Design, Innovation, and Discovery", Second Edition, 2005, John Wiley (New York, N. Y.).
- (6) Offshore Standard DNV-OS-F101, Submarine Pipeline Systems, Appendix E: Automated Ultrasonic Girth Weld Testing, Det Norske Veritas, October 2007.
- (7) 2007 ASME Boiler&Pressure Vessel Code, Section V – Nondestructive Examination, Article 14 – Examination System Qualification, The American Society of Mechanical Engineers, July 1, 2007, 2007 Edition. Addendum 2008a, BC03-1552.
- (8) MIL-HDBK-1823 (2009), Nondestructive Evaluation System Reliability Assessment, Department of Defense Handbook.

## Appendix 1 – Sample Summary Table for Estimation of AUT System Performance and Statistical Analysis

Date:  
Project No.:  
AUT Equipment:  
AUT Procedure:  
AUT Examiner:  
AUT Vendor:  
Location of Inspection:

### Data Table for Estimation of AUT Performance

[illegible]

FSH – Full screen height  
US/DS – Upstream (US) or Downstream (DS)