PVP2008-61612

A WEB-BASED DATA ANALYSIS METHODOLOGY FOR ESTIMATING RELIABILITY OF WELD FLAW DETECTION, LOCATION, AND SIZING (*)

Jeffrey T. Fong
Mathematical & Computational Sciences Division
National Institute of Standards & Tech. (**NIST**)
Gaithersburg, MD 20899-8910 USA
fong@nist.gov

Owen F. Hedden
Codes & Standards Consulting
3200 Overton Park W.
Fort Worth, TX 76109 USA
heddeno@sbcglobal.net

James J. Filliben
Statistical Engineering Division
National Institute of Standards & Technology (**NIST**)
Gaithersburg, MD 20899-8980 USA
filliben@nist.gov

N. Alan Heckert
Statistical Engineering Division
National Institute of Standards & Technology (**NIST**)
Gaithersburg, MD 20899-8980 USA
alan.heckert@nist.gov

ABSTRACT

Recent advances in computer technology, internet communication networks, and finite element modeling and analysis capability have made it feasible for engineers to accelerate the feedback loop between the field inspectors of a structure or component for critical flaws by nondestructive evaluation (NDE) and the office engineers who do the damage assessment and recommendations for field action to prevent failure. For example, field inspection data of critical flaws can be transmitted to the office instantly via the internet, and the office engineer with a computer database of equipment geometry, material properties, past loading/deformation histories, and potential future loadings, can process the NDE data as input to a damage assessment model to simulate the equipment performance under a variety of loading conditions until its failure. Results of such simulations can be combined with engineering judgment to produce a specific recommendation for field action, which can also be transmitted to the field by the internet. In this paper, we describe a web-based NDE data analysis methodology to estimate the reliability of weld flaw detection, location, and sizing by using a public-domain statistical data analysis software named DATAPLOT and a ten-step sensitivity analysis of NDE data from a two-level fractional factorial orthogonal experimental design. A numerical example using the 1968 ultrasonic examination data of weld seam in PVRC test block 251J, and the 1984 sectioning data of 251J containing 15 implanted flaws, is presented and discussed.

Keywords: Design of experiments; failure prediction; nondestructive evaluation; prediction intervals; reliability; statistical data analysis; structural engineering; tolerance intervals; ultrasonics; uncertainty estimation; web-based analysis; weld flaw detection.

1

1. INTRODUCTION

The need for a reliable method of detecting, locating, and sizing a crack by nondestructive evaluation (NDE) techniques such as ultrasonic testing (UT) or radiographic imaging as applied to a structure or component has been recognized by engineers and the public for years.

Since NDE techniques are fundamentally a measurement problem, its reliability was initially investigated in 1965 by the industry, as documented by Hedden [1], in a series of round-robin UT weld tests conducted by the Pressure Vessel Research Committee (PVRC) of the Welding Research Council. This testing program involved the use of 12 thick-section steel-plate weld specimens containing carefully designed flaws that were implanted and inspected by qualified teams using established ASME and other industry standards and procedures then-approved by PVRC.

Unfortunately, the results of the round-robin UT weld testing programs by a series of U.S., European, and Japanese teams over the next decade were negative, with very few exceptions. PVRC and the industry concluded that the then prevailing industry standards did not guarantee reliable and reproducible results. A similar conclusion was reached in 1976 as reported by Berger and Smith [2] in a study for the U.S. Department of Transportation (DOT) on the structural integrity of hundreds of girth welds of the Trans-Alaska Oil Pipeline:

"Defect dimensions can be determined with sufficient accuracy to be useful in the fracture mechanics analysis if the radiographs are made under carefully controlled conditions. If the radiographs are not made with close control, the accuracy of the defect sizes may not be sufficient to permit their use in establishing allowable defect sizes."

Since the report further stated that the field radiographs furnished by the pipeline owner were not made with close control, DOT concluded that there was not enough technical basis to grant the pipeline owner a waiver of then federal regulations on the non-acceptance of welds containing known defects. All defective welds except three buried under a river crossing were required to be repaired at a cost close to one hundred million 1976 dollars.

During the latter part of the 1970s and well into the 1980s, additional urgent need in the aerospace and nuclear power industries spurred rapid advances in NDE reliability not only in the United States (see, e.g., Lewis, et al [3], Johnson, et al [4], Adamonis and Hughes [5], Ruescher and Graber [6],

Yukawa [7-10], Fong et al [11-14]), but also in Europe (EU[15-17], Watkins, et al [18], etc.) and Japan (Saiga [19-22]). The industry realized that a new approach toward reliability was needed. The performance demonstration approach, modeled on the process specified by ASME for welding procedure development, was successfully developed and eventually was adopted by ASME and the industry in 1990. The NDE reliability problem was therefore successfully solved as shown in an address by the late Dr. Spencer Bush to a technical session of the International Symposium on Reliability of Reactor Pressure Components, IAEA, Stuttgart, Germany, March 21-25, 1983 (see a historical paper by Hedden [23] that was presented last year at the PVP San Antonio Conference):

". . . Reliability of flaw detection, sizing, and location represents a critical input in the overall assessment of nuclear systems and components comprising the pressure boundary. For example, a relatively benign flaw detected early in plant life can be evaluated by approved fracture mechanics techniques and permitted to remain indefinitely, subject to periodic monitoring, thus resulting in little or no perturbation in plant operation, plus generation of confidence in the safety authorities that the plant organization used "good" nondestructive examination procedures. "

During the last 25 years (1983-2008), advances in computer technology and internet communication were phenomenal to say the least. Application of NDE technology has also improved to the point that it is now feasible for field inspection data to be transmitted to the office via the internet, analyzed with sophisticated finite element models for damage assessment and remaining life prediction, with results of analysis and prediction transmitted back to the field within hours for prompt action (see, e.g., Fong, et al [24]).

A key to the success of this web-based, field-office-field data-transmission, and computer-modeling-and-analysis NDE capability, is the availability of an estimate of the reliability of the detection, location, and sizing of either a crack, or some measure of damage that will be used as input to a computer model to predict remaining life before failure. For example, if the field finds a crack at a certain location of a structure or component to be 2.0 in. long, it is essential that the field is capable of reporting that the crack length is $2.0 \pm$, say, 0.5 in., or, in another notation, 2.0 (0.5) in., with the uncertainty expressed as a 2-sided 95 % confidence interval, such that there is a 95 % probability that a single future NDE of that crack length is likely to be between 1.5 in and 2.5 in.

The purpose of this paper is, therefore, two-fold:
(1) To review the analysis methodology and the NDE reliability results as reported to PVRC in 1982-86 [11-14].
(2) To introduce a web-based design of experiments (DOE) approach to a field computer analysis of NDE data.
In Section 2, we review the details of the PVRC specimen 251J and its 15 implanted flaws. In Section 3, we re-visit the

1. Introduction (Continued)

analysis methodology and its implementation in a public-domain software named DATAPLOT [25-26]. In Section 4, we summarize the original results of Refs. [11-14] on the fabrication reliability of 3 types of flaws, namely, cross cracks, longitudinal cracks, and slags. Based on the 1968 PVRC ultrasonic testing (UT) data for detecting flaws in specimen 251J, as documented by Gillette and Smedley [27], White [28], Buchanan [29], and Hedden [1, 30-32], we summarize in Section 5 the original results of Refs. [11-14] on the reliability of 1968 UT data.

In Sections 6 and 7, we present a new approach to UT data analysis and uncertainty estimation by extending the results of the 1968 UT data to a new scenario where a series of 8 additional UT observations are assumed to be made in the field to justify a web-based analysis. This new approach uses a well-known technique in the statistics literature named the "Design of Experiments (DOE)." Significance and limitations of the DOE approach are discussed in Section 8, and a conclusion and references are given in Sections 9 and 10, respectively.

Since the work reported here was initiated in a 20-year period (1965-1985) using engineering units, we decided to use the same throughout the rest of the paper for ease of readability.

2. PVRC VALIDATION OF FLAW FABRICATION PROCEDURE USING SPECIMEN 251J (1981-86)

PVRC specimen 251J was fabricated by welding two 11-inch thick plates of ASTM A-533-65, Grade B, low-alloy steel (80,000 psi tensile strength), using the submerged arc welding process. The dimensions of the specimen and a sketch of the relative locations of the 15 implanted flaws are given in Figs. 1 and 2. Of the 15 flaws, 5 were longitudinal cracks (C, D, H, L, M), 5 cross cracks (A, E, I, J. N), 3 long slag inclusions (B, G, O), and 2 short slag inclusions (F, K). The actual shape and location of each flaw after sectioning, as first reported by Yukawa [7], were quite complicated, as illustrated in a graphical representation of one of the 15 flaws (Fig. 3). A comparison of the intended flaw location vs. sectioning data as projected on the y-z plane is shown in Fig. 4. Defining an amplification factor (AF) as the ratio of the actual maximum dimension of an implanted flaw to the intended maximum, we obtained the following list of AF values for each of the 15 intended flaws:

Longitud. Cracks: AF (C, D, H, L, M) = (1.575, 1.650, 2.750, 2.375, 2.200).

Cross Cracks: AF (A, E, I, J. N) = (2.133, 2.667, 2.467, 2.000, 1.733).

Slags: AF (B, G, O; F, K) = (1.000, 1.300, 1.225; 1.467, 1.267).
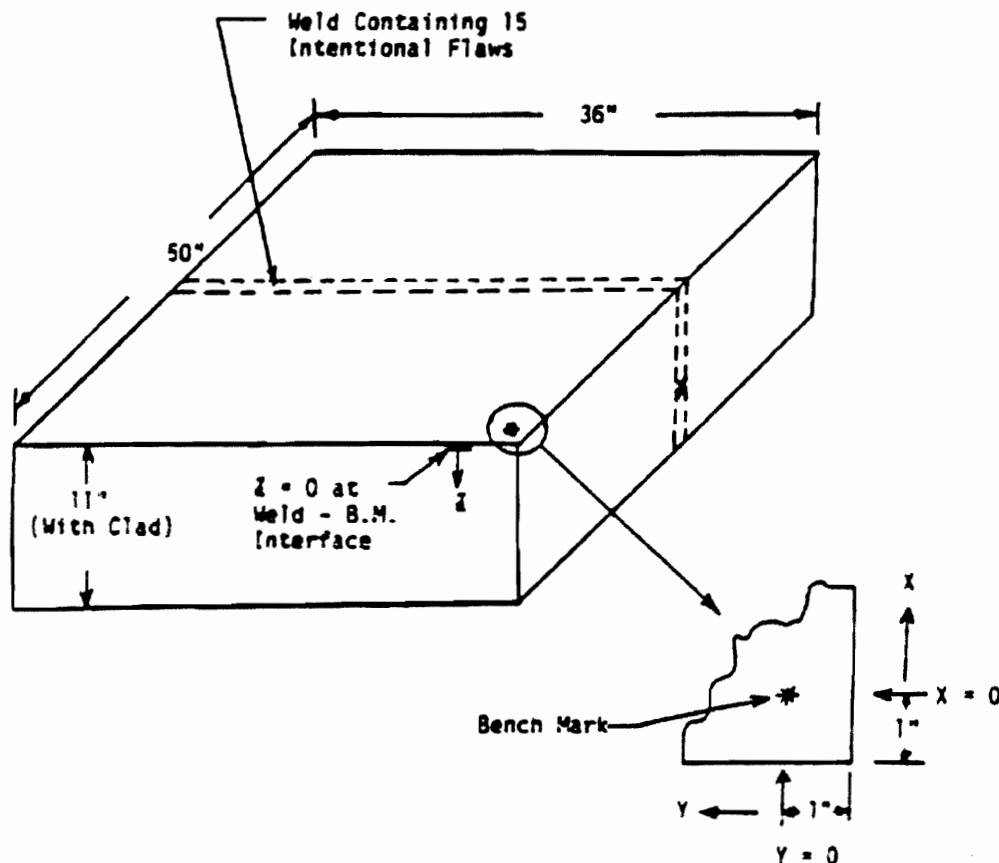
Fig. 1. Dimensions of PVRC Specimen 251J. After Yukawa [7, 8].
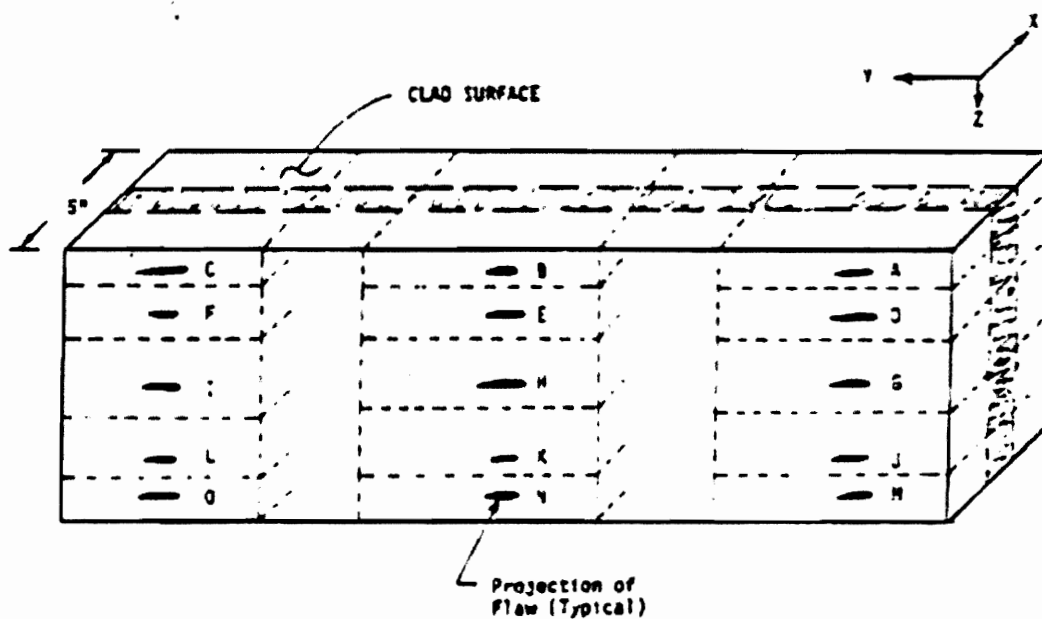
3

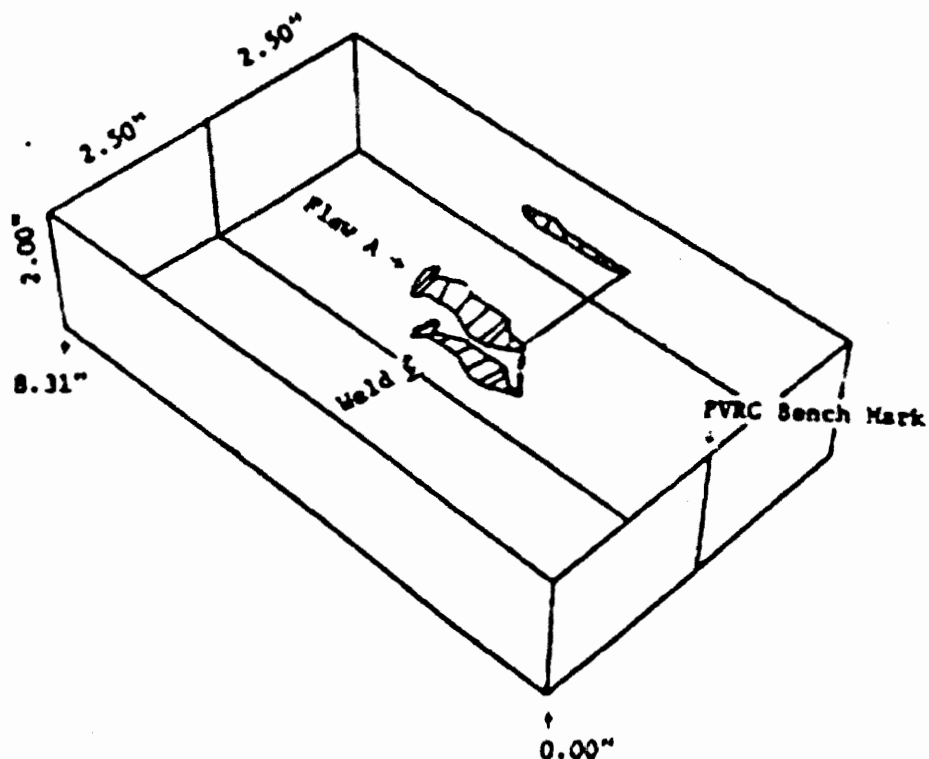Fig. 2. Principal Sectioning Cuts of PVRC Specimen 251-J. After Yukawa [7, 8].

Fig. 3. A Three-dimensional Computer-aided Graphical Representation
of Flaw A based on its Projected Outline as reported by Yukawa [7].

4

[PVRC| INT DP1

PVRC 251J: INTENDED FLAWS (SOLID) VS ACTUAL (DOTTED)

.... Sub-Block 1 .... Sub-Block 2 .... Sub-Block 3 ....

ELEVATION (Y-HORIZ., Z-VERT., UNITS IN INCHES)

Fig. 4. Comparison of intended (solid lines) vs. actual (dotted lines) locations of 15 flaws as shown in Fig. 2. Note: **CC** = cross cracks (A, E, I, J, N). **LC** = longitudinal cracks (C, D, H, L, M). The rest are slags, 3 long (B, G, O) and 2 short (F, K) ones.

Normal Probability Plot Correlation Coefficient (PPCC) Test for Normality

Application No. 2:   Reliability of PVRC 15-Flaw Fabrication (Fong-Hedden, 1986)

Normal PPCC Test Statistic Value = 0.985 (n = 15)

Since 5% Point of Normal PPCC = 0.937 < 0.985 , test statistic < 1.0

Conclusion: Normality Hypothesis Accepted.

tint215.dp

< via Dataplot >

Normal N (0,1) Order Statistic Medians

Fig. 5.   Probability plot for testing a normal distribution fit for 15 data points of Flaw Size Amplification Factor using a **P**robability **P**lot **C**orrelation **C**oefficient (**PPCC**) test for normality (Filliben [35]) as implemented in DATAPLOT [25, 26].

3. A STATISTICAL ANALYSIS METHODOLOGY FOR ESTIMATING FLAW FABRICATION RELIABILITY

As outlined in Ref. [12], the analysis of the 15 maximum flaw dimension amplification factors, AF, as displayed in the last sect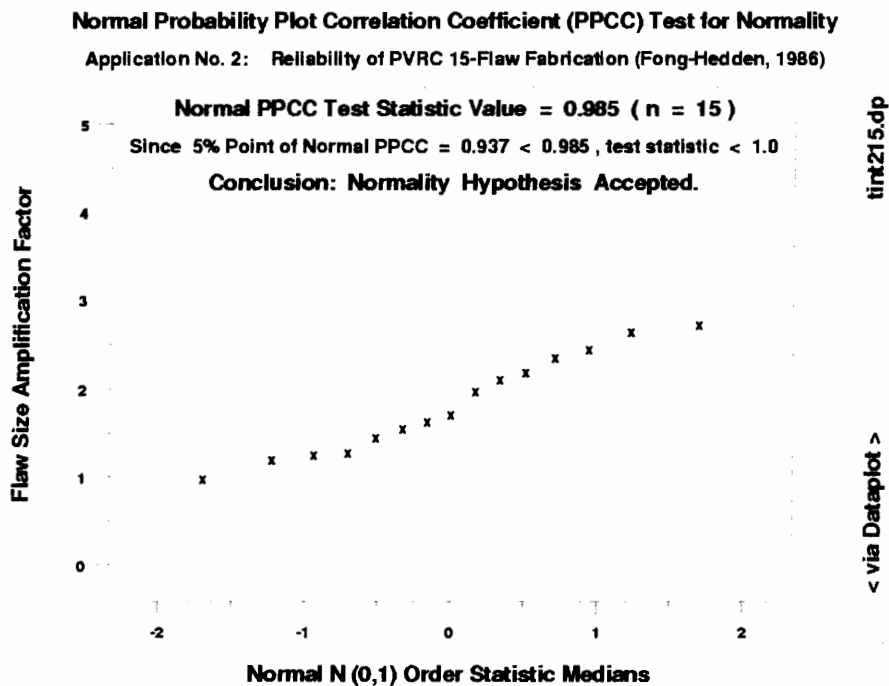ion, consisted of five distinct steps. For brevity, we only report the results of the first three steps, which are elementary (see, e.g., Nelson, et al [33]). Step 4 on analysis of variance and step 5 on tolerance limit intervals will be explained in detail. The first three steps are:

Step 1 Univariate Analysis   In this analysis, we made the naive assumption that the 15 values of AF were equally representative of the quality of the flaw fabrication procedure. The sample average (M) was 1.854, and the sample standard deviation (S) was 0.560.

Step 2 Test for Normality   Having calculated M and S, we needed to answer two questions, namely, (a) was the distribution normal? and (b) was the distribution homogeneous? Using a call routine in DATAPLOT [25-26], which is based on the Tukey's Lambda Test for families of symmetric distributions (Filliben [34-35]), we found that the probability plot correlation coefficient (PPCC) for the distribution of the 15-data set to be normal was 0.985, which was close enough to unity to justify the normality assumption (see Fig. 5).

Step 3 Test for Homogeneity   To examine whether the 15-data set was reasonably homogeneous even though it involved four types of flaws, we used the "boxplot" routine in DATAPLOT [25-26] to obtain a display of the data and to make a visual observation that there appeared to be an effect due to flaw types, and that the fabrication procedure appeared to be "good" for both types of slag inclusions, but "not so good" for the longitudinal and cross cracks. This observation led us to the next step, i.e., the use of the analysis of variance to determine the effect of the flaw types. For more details of the above three steps, we refer the reader to the original report by Fong [12].

Step 4 Analysis of Variance (ANOVA)   One of the best known and powerful data analysis techniques in statistics is ANOVA, which originated in the 1930s with Daniels [36], and was later discussed by many including Eisenhart [37], Crump [38], Hendricks [39], and Mandel [40-41]. In this expository paper, we follow Draper and Smith [42, pp. 423-454] and use the "anova" routine in DATAPLOT [25] to obtain not only the two earlier estimates, M (=1.854) and S (0.560), but also a new estimate of the standard deviation due to replication (S_R) equal to 0.382. To compute the standard deviation due to flaw types (S_I), the following formula (see, e.g., Mandel [41]) may be used if the number of replicas, to be denoted by m, in each subgroup is contant:

$$(S_I)^2 = (S)^2 - (S_R)^2 / m. \qquad (1)$$

For our 15-data set, m varied from 5 for the two types of cracks to 3 for long slags and 2 for short slags. If we denote the numbe of types by t ($= 4$ in this case), and $m_1, m_2, ..., m_t$, the number of data in each type, we can introduce the notion of an "equivalent number of replicas," to be denoted by m_m, as the integer closest to and greater than the average of all the m's. In other words,

$$m_m = \text{INT} \{ (m_1 + m_2 + ... + m_t) / t + 0.5 \}. \qquad (2)$$

For our 15-data set, $t = 4$, $m_1 = m_2 = 5$, $m_3 = 3$, and $m_4 = 2$. Using eq. (2), we calculated that $m_m = 4$. Using eq. (1), and the values of m_m, S, S_R, we found that $S_I = 0.5264$, which was different from either S or S_R. We concluded that there was indeed a flaw type effect and the 15-data set was not homogeneous.

To incorporate the effect of the flaw types, as represented by S_I, into the overall estimate of the variance of the non-homogeneous 15-data set, we can combine the two standard deviations, S_R, and S_I, into a revised standard deviation, to be denoted by S_2, as follows:

$$(S_2)^2 = (S_R)^2 + (S_I)^2. \qquad (3)$$

For the 15-data set under consideration, S_2 was found to be 0.650, which was about 16% higher than S ($= 0.560$) of the univariate analysis. To continue to the next step, which is to set us a numerical criterion whether to accept or reject a specific procedure such as the flaw fabrication or ultrasonic testing, we shall use S_2 instead of S to account for the effect due to flaw types.

Step 5. Tolerance Intervals as Acceptance Criterion   Having estimated the global mean, M, and the revised standard deviation, S_2, of the 15-data set, we could use a formula (see, e.g., Nelson, et al [33, pp. 178-180, eq.5.3.2]) and the well-known tables of values of the t-distribution to calculate the so-called "prediction intervals," which indicate where a *single* future observation is likely to be. But we need more than that if we are interested in establishing a criterion for accepting or rejecting a procedure. The notion we need is the "tolerance intervals," which indicate where a proportion, P, or "coverage", of the population is likely to be.

Let us make a simplifying assumption that one can estimate the reliability of flaw fabrication solely on a single metric, namely, the amplification factor, AF. Following Proschan [43] and using the tables of tolerance factors K, first tabulated by Natrella [44], later reproduced by Beyer [45], and also reproduced here in the end of this paper as Tables 1 and 2, for various sample sizes n ($= 2, 3, ...,$ infinity) and five discrete values of P (0.75, 0.90, 0.95, 0.99, 0.999), we can estimate the reliability of the flaw fabrigation procedure as the probability, γ, between 75% and 99% , such that at least a proportion P will be included between $M - K S_2$ and $M + K S_2$. Since Tables 1 and 2 are for discrete values of γ, we used a fit routine to obtain a continuous set of numbers for a specific sample size and coverage as illustrated in Figs. 6 and 7.

6

Normal Tolerance Limit Factors (n = 15, Coverage P = 90 %)

Ref.: Natrella, M. G., Experimental Statistics, NBS Handbook 91 (1966).

Let y = Tolerance Factor (see Tables 1 & 2).
Let x = Reliability Gamma (see Tables 1 & 2).
Let the model for a 4-parameter set be given by:

$$y = (a0 + a1*x + a2*x**2) / (1 + b1*x)$$

×2.945

2.480

2.278

1.994

**Factors for Two-Sided Tolerance Limits
for Normal Distributions
Coverage P = 90%; Sample Size n = 15
(Tables 1 & 2 from Natrella's Book, pp. T-10 & T-11)**

Reliability or Probability γ

Tolerance Factor

tint215.dp

< via Dataplot >
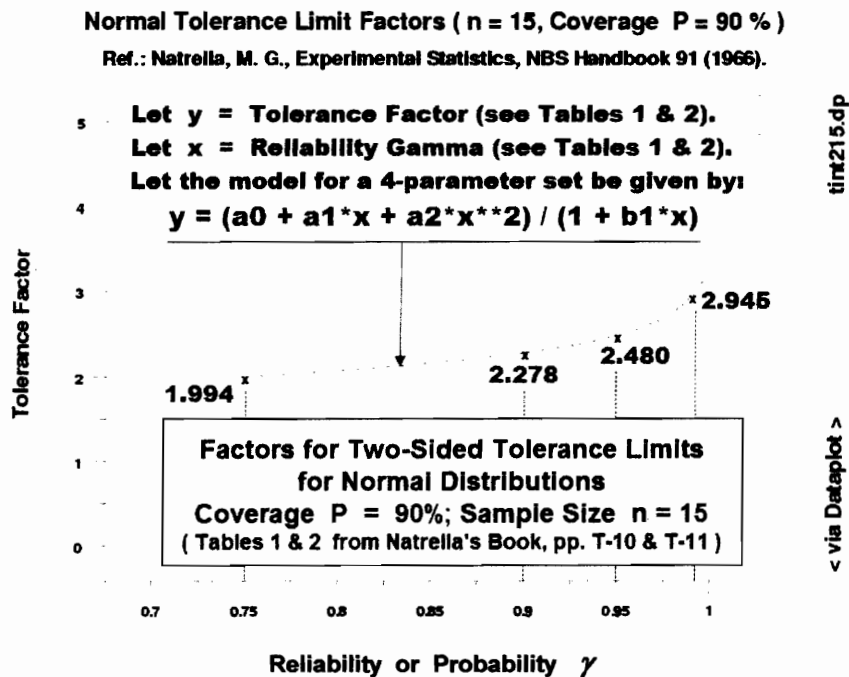
0.7   0.75   0.8   0.85   0.9   0.95   1

Fig. 6. A plot of **TFN**$_{90,15}$, the two-sided **T**olerance limit **F**actors for **N**ormal distributions at coverage $P = 90\%$ and sample size $n = 15$, vs. γ, the reliability or probability that at least a proportion P of the distribution will be included between $X \pm$ **TFN**$_{90,15} * s$, where X and s are estimates of the mean and the standard deviation computed from a sample size n.

Normal Tolerance Limit Factors (n = 15, Coverage P = 99 %)

Ref.: Natrella, M. G., Experimental Statistics, NBS Handbook 91 (1966).

tint215.dp

Reliability

< via Dataplot >

For P = 90%	
Gamma	T.F.
x	y
0.75	1.994
0.90	2.278
0.95	2.480
0.99	2.945

For P = 99%	
Gamma	T.F.
x	y
0.75	3.118
0.90	3.562
0.95	3.878
0.99	4.605

For P = 95%	
Gamma	T.F.
x	y
0.75	2.375
0.90	2.713
0.95	2.954
0.99	3.507

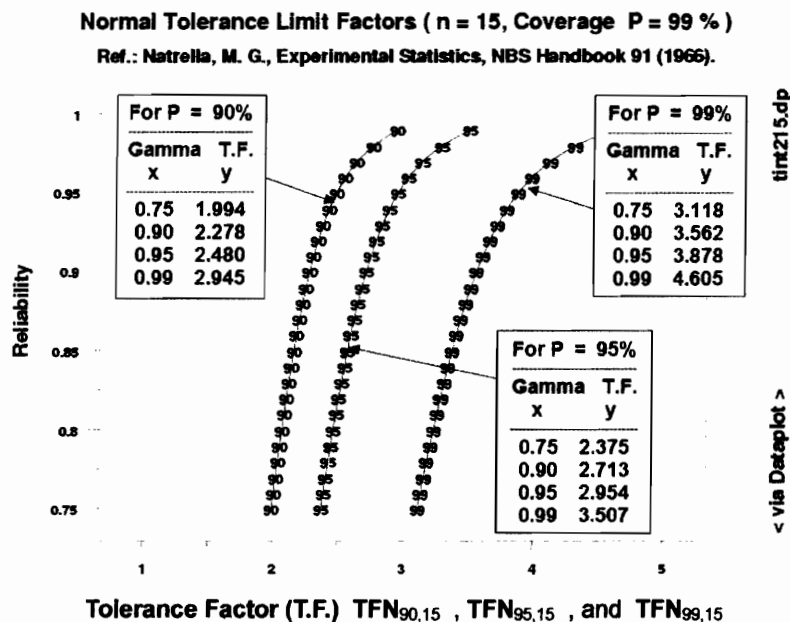Tolerance Factor (T.F.) TFN$_{90,15}$, TFN$_{95,15}$, and TFN$_{99,15}$

Fig. 7. A plot of gamma, γ, the reliability or probability that at least a proportion P of the distribution will be included between $X \pm$ **TFN**$_{P,n} * s$, where X and s are estimates of the mean and the standard deviation computed from a sample size n, vs. Tolerance Factor (T.F.), or, **TFN**$_{P,n}$, the 2-sided **T**olerance limit **F**actors for **N**ormal distributions at coverage P and sample size n.

7

4. PVRC-1968 Flaw Fabrication Reliability

The plots of the probability γ vs. K, or $TFN_{P,n}$, the two-sided **T**olerance limit **F**actors for **N**ormal distribution at coverage P and sample size n, as illustrated in Fig. 7 for $P =$ 0.90, 0.95, 0.99, and $n = 15$, can now be used to calculate the tolerance limits, $M - KS$ and $M + KS$, for an estimated mean M and standard deviation S. In this case, we interpret the probability γ as the reliability of a specific procedure under consideration for a coverage P using data from a sample n.

We are now ready to assign reliability to a total of 5 cases in the 1968 PVRC weld flaw fabrication procedure.

Case 1   Consider all 15 flaws as a group.. No ANOVA is done. $M = 1.854$. $S = 0.560$. Three plots of reliability γ vs. AF are given in Fig. 8 for coverages $P = 0.90, 0.95$, and 0.99. We purposely include this case to allow us a comparison with Case 5, where the flaw type effect is accounted for using ANOVA. For a 90% coverage and 95% reliability, max $AF = 3.243$.

Case 2   Consider 5 cross cracks only. No ANOVA is done. $M = 2.200$. $S = 0.372$. Plots of γ vs. AF are given in Fig. 9. For a 90% coverage and 95% reliability, max $AF = 3.788$.

Case 3   5 longitudinal cracks only. No ANOVA is done. $M = 2.110$. $S = 0.496$. Plots of γ vs. AF are given in Fig. 10. For a 90% coverage and 95% reliability, max $AF = 4.232$.
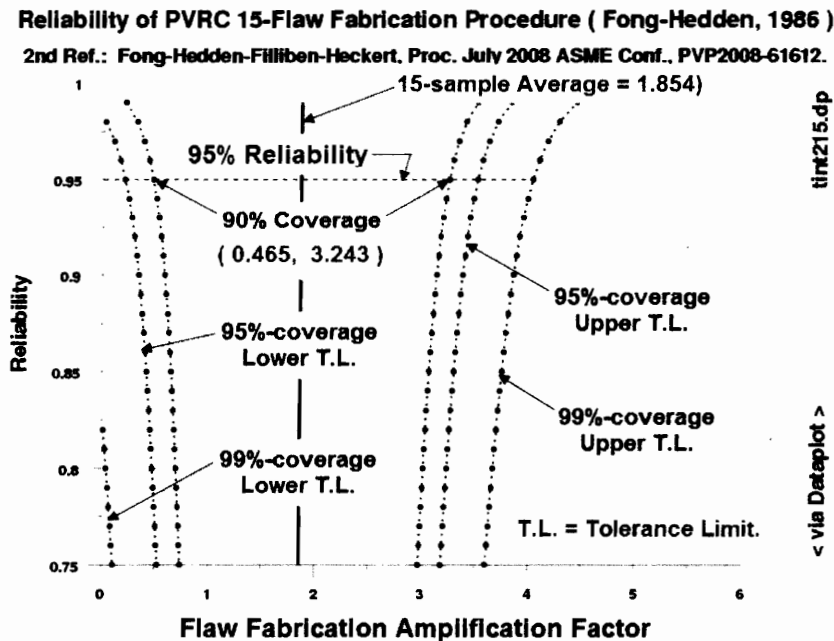
Case 4   5 slag inclusions only. No ANOVA is done. $M = 1.252$. $S = 0.168$. Plots of γ vs. AF are given in Fig. 11. For a 90% coverage and 95% reliability, max $AF = 1.970$.

Case 5   Consider all 15 flaws as a group with the flaw type effect determined by ANOVA. In this case, $M = 1.854$, $S = S_2 = 0.650$. Plots of γ vs. AF are given in Fig. 12. For a 90% coverage and 95% reliability, max $AF = 3.466$, which is larger than 3.243 given in Case 1.

For a coverage of 90%, the following summarizes the results:

Case	Type of Flaws	Upper Limit of AF	Fig. No.
5	All flaw types	3.466	12
2	Cross Cracks	3.788	9
3	Longitudinal Cracks	4.232	10
4	Slag Inclusions	1.970	11

The above allows us to conclude that if the acceptance criterion is 2.000, the fabrication procedure passes only for the slags, but fails for all flaw types, cross cracks, or longitudinal cracks.

Reliability of PVRC 15-Flaw Fabrication Procedure (Fong-Hedden, 1986)

2nd Ref.: Fong-Hedden-Filliben-Heckert. Proc. July 2008 ASME Conf., PVP2008-61612.

Fig. 8. A plot of gamma, γ, the reliability of flaw fabrication as defined by the two-sided tolerance limit intervals, vs. the flaw fabrication amplification factor for 3 coverages, $P = 90\%, 95\%$, and 99%, and a sample size $n = 15$. For example, for a coverage of 90%, the flaw fabrication amplification factor estimated from 15 flaws with a certified normal distribution and without ANOVA for flaw types, is given by a mean of 1.854 and a tolerance interval of (0.465, 3.243), or alternatively, 1.854 (1.389), with 95% reliability. As the coverage increases, so does the tolerance interval for the fabrication amplification factor.

8

4. *PVRC-1968 Flaw Fabrication Reliability (continued)*

Reliability of PVRC 5-Cross Crack Fabrication (Fong-Hedden, 1986)

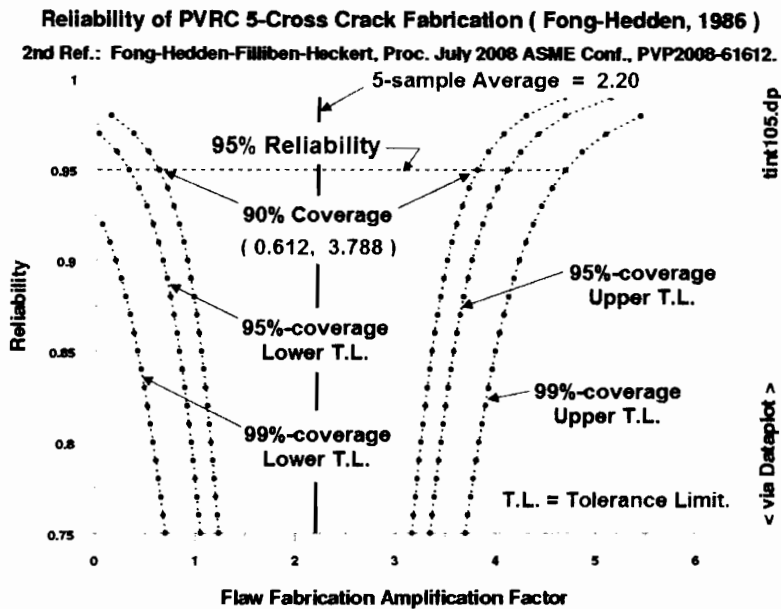2nd Ref.: Fong-Hedden-Filliben-Heckert, Proc. July 2008 ASME Conf., PVP2008-61612.

Fig. 9. A plot of gamma, γ, the reliability of flaw fabrication, vs. the **cross crack** fabrication amplification factor for 3 coverages, P = 90%, 95%, and 99%, and a sample size n = 5. For example, for a coverage of 90%, the **cross crack** fabrication amplification factor estimated from 5 such cracks with a certified normal distribution, is given by 2.20 (1.588), with 95% reliability. For brevity, plots of γ vs. **TFN**$_{P, n}$ for P = 90%, 95%, 99%, and n = 5, similar to Figs. 6 and 7, are omitted.

Reliab. of PVRC 5-Longitudinal Crack Fabrication (Fong-Hedden, 1986)

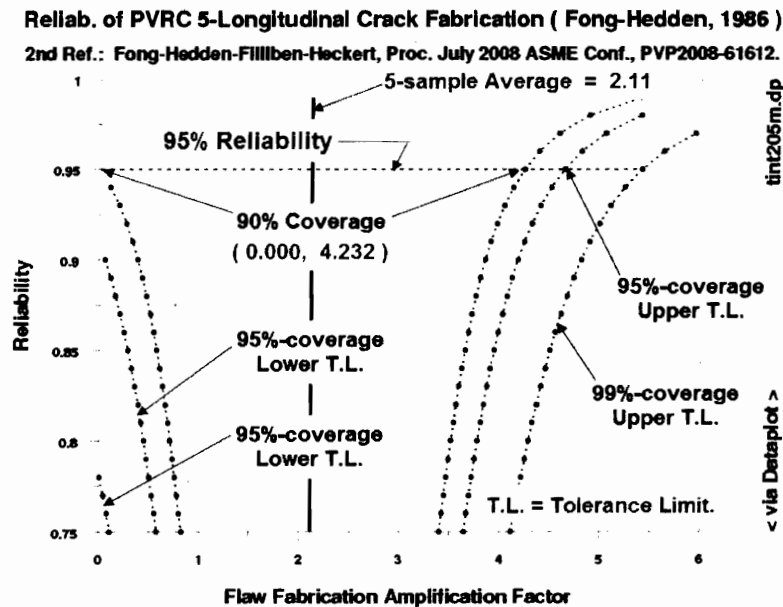2nd Ref.: Fong-Hedden-Filliben-Heckert, Proc. July 2008 ASME Conf., PVP2008-61612.

Fig. 10. A plot of gamma, γ, the reliability of flaw fabrication, vs. the **longitudinal crack** fabrication amplification factor for 3 coverages, P = 90%, 95%, and 99%, and a sample size n = 5. For example, for a coverage of 90%, the **longitudinal crack** fabrication amplification factor estimated from 5 such cracks with a certified normal distribution, is given by 2.11 (2.122), with 95% reliability. Again for brevity, plots of γ vs. **TFN**$_{P, n}$ for P = 90%, 95%, 99%, and n = 5, are omitted.
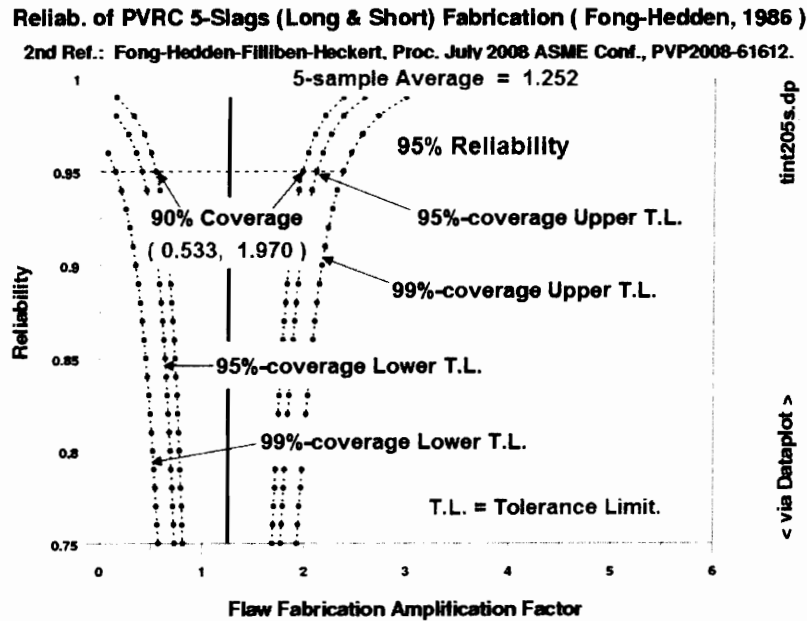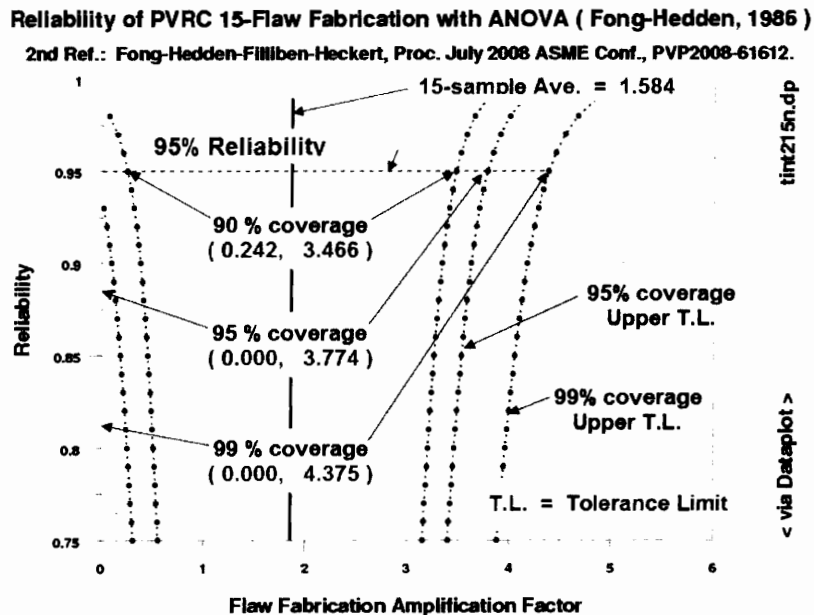
Fig. 11. A plot of gamma, γ, the reliability of flaw fabrication, vs. the **slag** fabrication amplification factor for 3 coverages, P = 90%, 95%, and 99%, and a sample size $n = 5$. For example, for a coverage of 90%, the **slag** fabrication amplification factor estimated from 5 such flaws with a certified normal distribution, is given by 1.252 (0.718), with 95% reliability. Again for brevity, plots of γ vs. **TFN**$_{P,n}$ for P = 90%, 95%, 99%, and $n = 5$, are omitted.

Fig. 12. A plot of gamma, γ, the reliability of flaw fabrication, vs. the flaw fabrication amplification factor with ANOVA for flaw types and for 3 coverages, P = 90%, 95%, and 99%, and a sample size $n = 15$. For example, for P = 90%, the flaw fabrication amplification factor estimated from 15 flaws with a certified normal distribution and with ANOVA for flaw types, is 1.854 (1.612), with 95% reliability. Note that with ANOVA for flaw types, the half-interval increases from 1.389 to 1.612.

10

5. ULTRASONICS-1968 DETECTION RELIABILITY

One of the main difficulties in assessing the reliability of a complex procedure such as the nondestructive evaluation (NDE) technique of detecting, locating, and sizing a flaw is the requirement for a large quantity of data normally unavailable because of a time or cost constraint, or both. The lack of an a priori knowledge of the distribution or variability of many instrument and human factors creates a barrier for producing a credible reliability analysis based on a small amount of round robin data.

On the other hand, if we can show that the underlying distribution of a "response variable" is almost or close to normal, a combination of the techniques of ANOVA and tolerance factors can be applied even though the quantity of data is "small." An example of this approach has been presented earlier in assessing the reliability of a flaw fabrication procedure. In this section, we shall present a similar analysis of the 3-team ultrasonic testing UT-1968 data by first choosing a response variable, then testing for normality, and finally applying the ANOVA technique to evaluate the "between-team," "flaw-type," and the so-called DAC effects on the 3-team UT-1968 data. A complete listing of the 3-team UT-1968 data is given in Tables 3, 4, and 5 at the end of this paper.

The response variable we choose to work with is the "detection threshold" with the unit of "inch" for flaw length, to be denoted by DT, at which new and distinct flaws are detected. We define this variable to have the further property that the height of its histogram be zero at two values of DT, namely, $DT = 0$, and $DT = DT_{max}$, where DT_{max} is some large number. The central question we ask is whether the distribution of the varialbe DT using data between the two ends, $DT = 0$, and $DT = DT_{max}$, is close to normal. If that were the case, we can apply the analysis methodology described in Section 3 to yield some information on the reliability of the UT-1968 detection capability. If not, we need to develop an alternative methodology such as non-parametric (or distribution-free) prediction or tolerance intervals as described by Nelson, et al [33, pp. 181-187].

In Table 3 and Fig. 13, we present the UT-1968 data of Team A as corrected by Hedden [1, 30, 31]. Note that Team A produced a total of 18 indications, even though specimen 251J was known to contain only 15 flaws.

$[PVRC]$ $UT68AY.DP$

TEAM "A" – UT DATA (HEDDEN, 84-03-07), 1968 PROCEDURE

ELEVATION (Y-HORIZ., Z-VERT., UNITS IN INCHES)
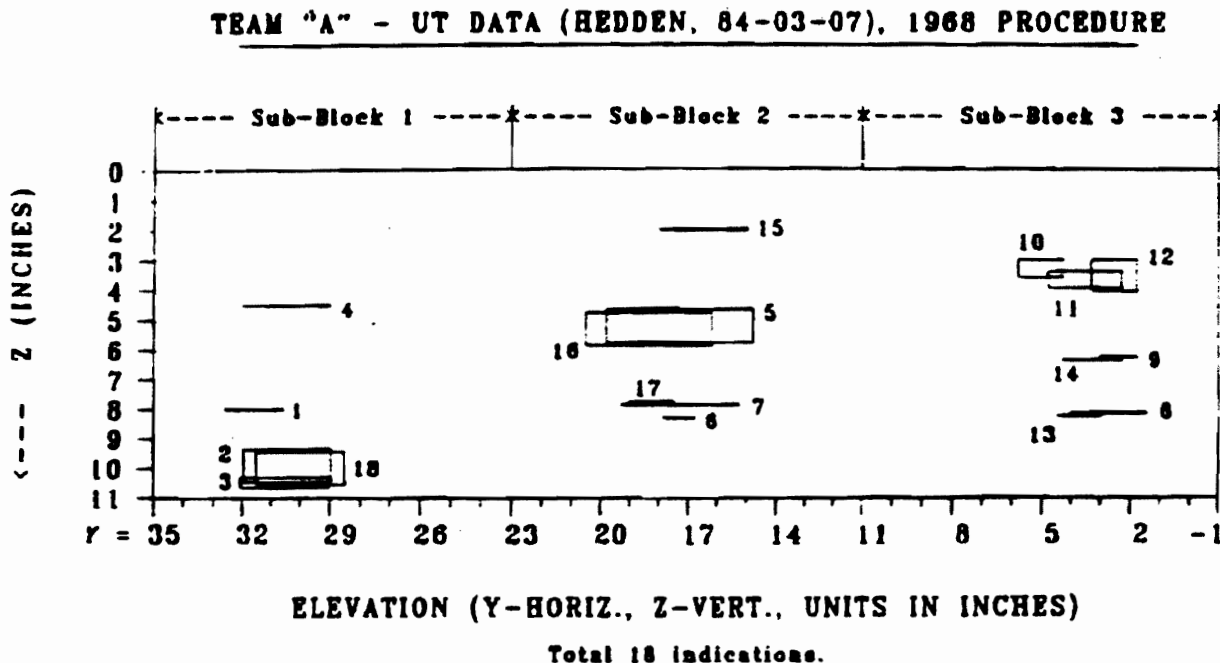
Total 18 indications.

Fig. 13. Elevation of PVRC Specimen 251J showing the location and sizing of all 18 ultrasonic indications (see Table 3) of the flaws first identified by **Team A** in 1968 [27,28], analyzed by Buchanan [29] in 1976, and reproduced by Hedden [1, 30, 31] in 1981 & 1984.

11

5. Ultrasonics-1968 Detection Reliability (Continued)

For each of the 18 indications reported by Team A, we conducted a 3-step analysis as described below:

Step 1   Center-to-Center Distance (*CCD*) Computation

For this step, we converted all box-type data for the 15 implanted flaws into centroid and half-length data to facilitate the computation of the *CCD* between an indication and each of the 15 implanted flaws. This yielded 15 *CCD's*, which we then ordered in an ascending sequence for a comparison with a set of threshold values.

Step 2   Identification of Flaws Detected vs. Threshold

For this step, we defined an ascending sequence of flaw detection thresholds between a minimum of ¼ inch and a maximum of 3.0 in. in increments of ¼ inch. For each value of that detection threshold and each indication, we identified the flaw or flaws whose CCD's were equal or less than the threshold. A new sequence of positive integers was then found to characterize the detection capability of that indication as a function of the detection threshold sequence.

Step 3   Graphical Representation of Each Indication

For this step, we plotted the number of flaws detected vs. the detection threshold and display (a) the name of the nearest flaw detected, (b) the CCD, (c) the ratio of the maximum dimension of actual flaw to that of the detected indication (also known as the size factor), and (d) the maximum DAC value reported for that indication. An example of this display is given in Fig 14 for Team A's indication #1, Fig. 15 for Team A's indication #12, Figs. 16-17 for Team B's #4, and Fig.

Fig. 14. Result of an automatic detection data analysis for **Team A Indication #1** with a variable detection threshold to parametrize the decision-making process.
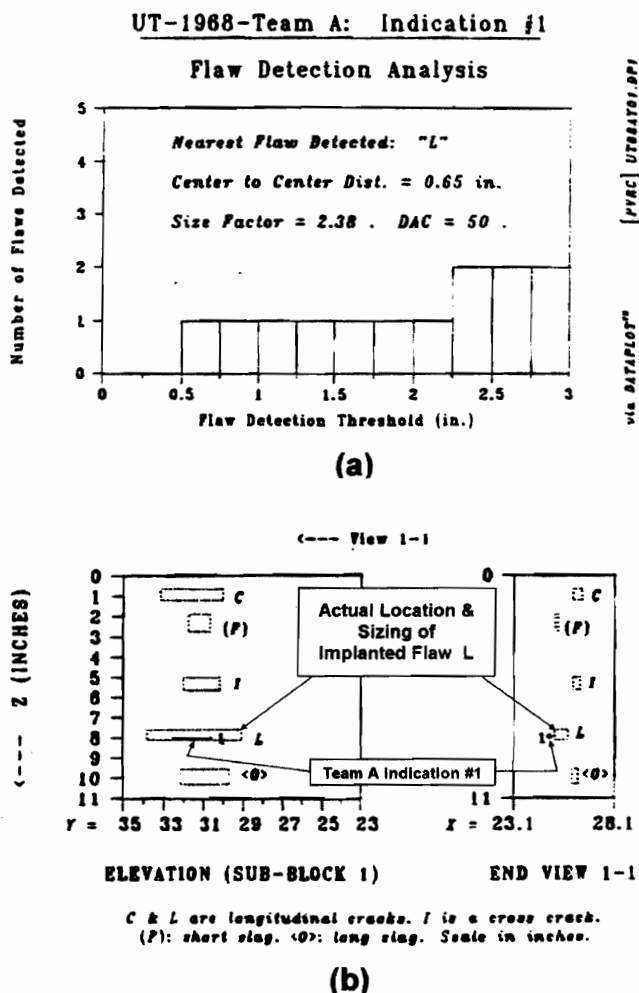

Fig. 15. Result of an automatic detection data analysis for **Team A Indication #12** with a variable detection threshold to parametrize the decision-making process.
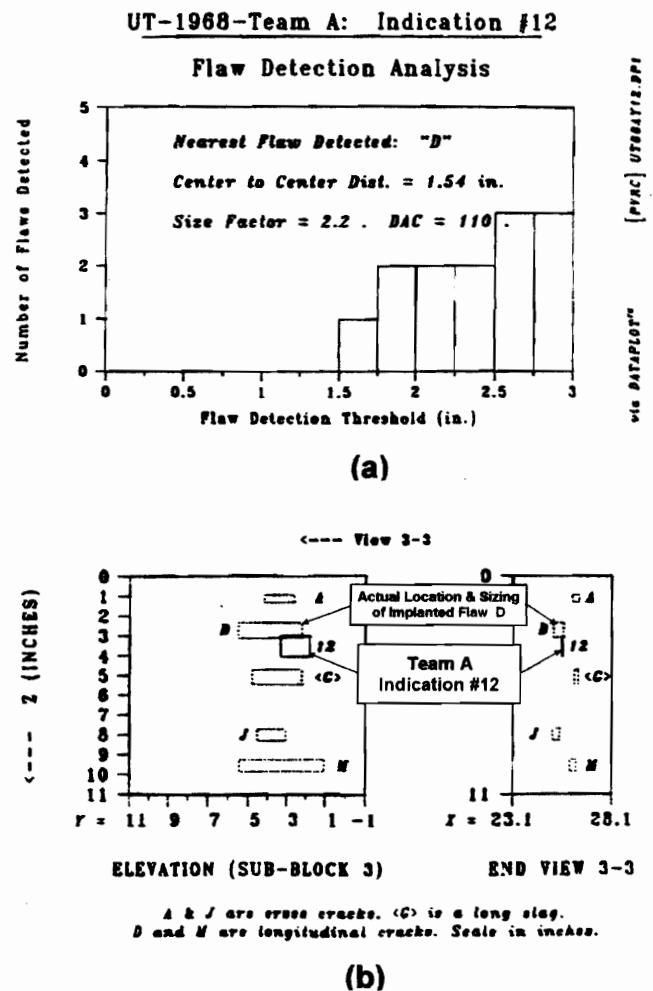
12

5. Ultrasonics-1968 Detection Reliability (Continued)

18 for Team C with the display for one of its 14 indications omitted for brevity.

By collecting 18 such displays for Team A, 4 for Team B, and 14 for Team C, we can assess the detection capability of each team by defining two probabilities and one activity metric, each as a function of the detection threshold, as described below:

(1) True-Call Probability (TCP). For each detection threshold and each indication, the analysis so far allowed us to answer the question whether any flaw has been identified. If the answer is "yes", we assign a "1" to the indication as a "true-call." Otherwise, we assign a "0" to show that it is a "false-call." The TCP is defined as the ratio equal to the total number of true calls registered divided by the total number of indications reported. If an indication identifies more than one flaw, as in the case of Fig. 14 (Team A, Indication #1) for threshold equal to or greater than 2.25, the true call count is still 1. Note that this definition of TCP allowed us to define a "false-call" probability (FCP) by the expression FCP = 1 - TCP. Results of the TCP's of Teams A, B, and C are given in Figs. 19-21, respectively. This led us to conclude that a 2-inch threshold is

enough for all 3 teams to claim a 100% TCP or 0% FCP.

(2) Flaw Detection Probability (FDP). Among the true calls, it happened that some of the flaws were identified more than once. The purpose of introducing FDP was to eliminate the duplicate calls by counting exactly how many flaws the team had found. Thus we defined FDP to be the ratio between the total number of distinct flaws detected and the total number of implanted flaws. Results of the FDP's of Teams A, B, and C, are again given in Figs. 19-21, respectively. A visual inspection of the three figures shows that both Teams A and C did well on both TCP and FDP, but Team B did well on TCP but failed on FDP.

(3) Detection Activity Index (DAI). The two probabilities defined above measured the "quality" or "efficiency" of a team's detection capability, and it is useful to define a third metric simply to measure the detection activity of each team as the ratio equal to the total number of true calls divided by the total number of implanted flaws. Results of the DAI's for three types of flaws reported by Teams A, B, and C, are given in Figs. 22-24, respectively. Such graphical representation of the UT data allowed us to apply the technique of ANOVA to determine the flaw type effect in the capability of each team.

Fig. 16. Elevation of PVRC Specimen 251J showing the location and sizing of all 4 ultrasonic indications (see Table 4) of the flaws first identified by **Team B** in 1968 [27,28], analyzed by Buchanan [29] in 1976, and reproduced by Hedden [1, 30,31] in 1981 & 1984.

13

UT-1968-Team B: Indication #4

Flaw Detection Analysis

Nearest Flaw Detected: "0"

Center to Center Dist. = 1.05 in.

Size Factor = 0.63 . DAC = 30 .

Number of Flaws Detected (y-axis): 0 to 5

Flaw Detection Threshold (in.) (x-axis): 0, 0.5, 1, 1.5, 2, 2.5, 3

[PVRC] UT68B04.DP1

via DATAPLOT"
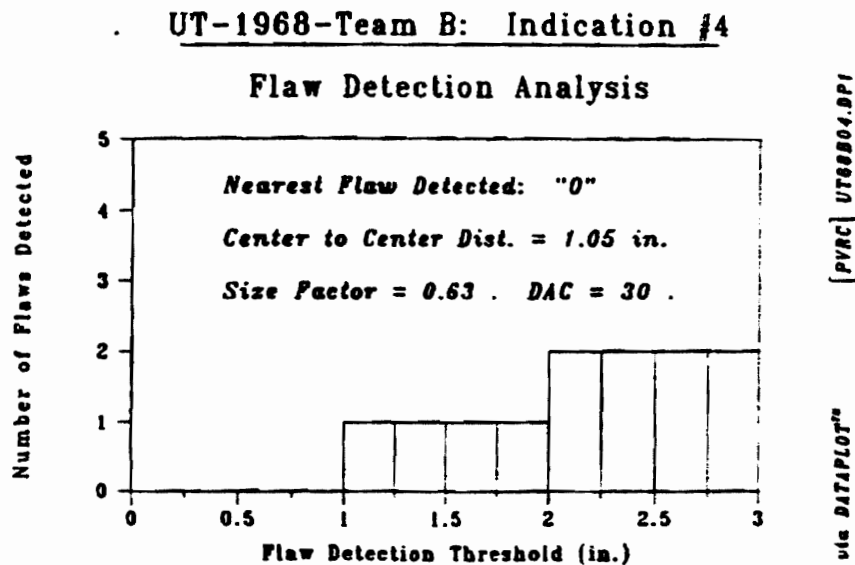
Fig. 17. Result of an automatic detection data analysis for **Team B Indication #4** with a variable detection threshold to parametrize the decision-making process. For brevity, we omit the actual vs. detected flaw location plot.

[PVRC] UT68C.DP

TEAM "C" — UT DATA (Uncorrected), 1968 Procedure

|----- Sub-Block 1 ----*----- Sub-Block 2 ----*----- Sub-Block 3 ----|

Z (INCHES): 0 to 11

• 11 , 1 , 6

7 , •12 , 4

2 , 8 , •13

, 5 ,

10 , •14 , 3

r = 35, 32, 29, 26, 23, 20, 17, 14, 11, 8, 5, 2, -1

ELEVATION (Y-HORIZ., Z-VERT., UNITS IN INCHES)
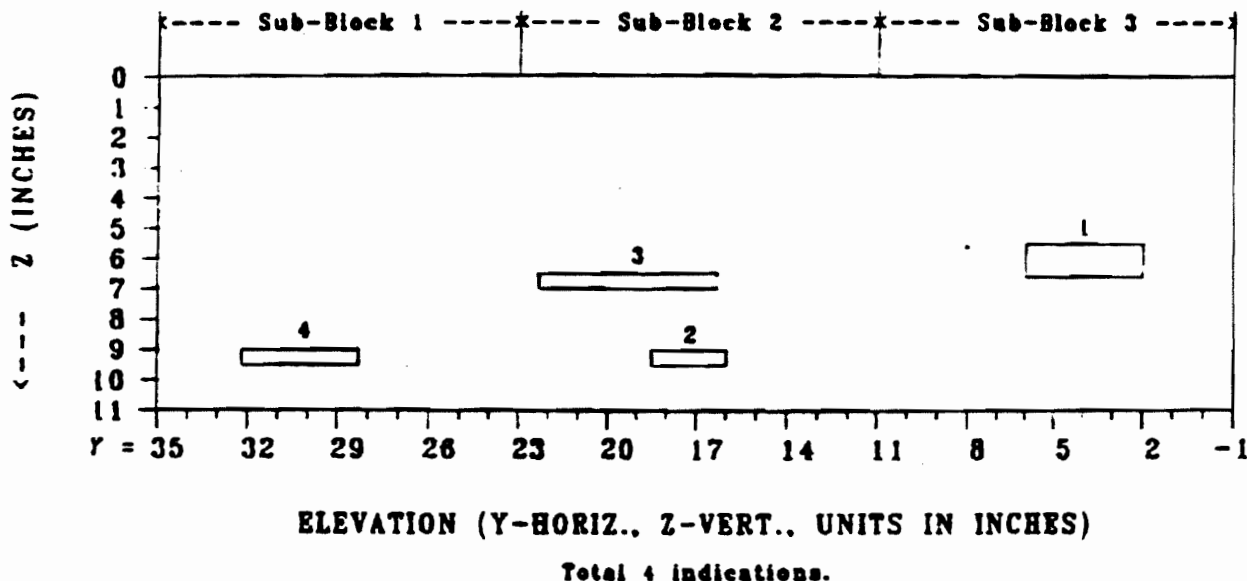
Total 14 indications.
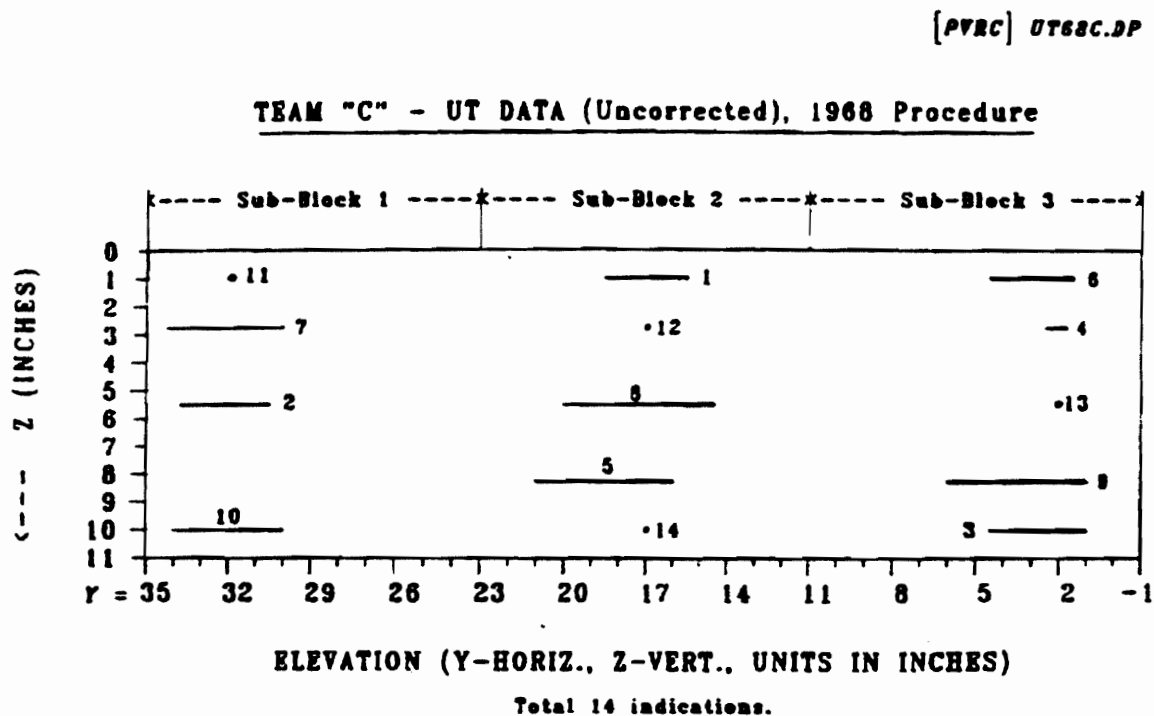
Fig. 18. Elevation of PVRC Specimen 251J showing the location and sizing of all 14 ultrasonic indications (see Table 5) of the flaws first identified by **Team C** in 1968 [27,28], analyzed by Buchanan [29] in 1976, and reproduced by Hedden [1, 30,31] in 1981 & 1984.

14

UT Flaw Detection Probability

PVRC-251J (Hedden, 84-03-07)

Team: A. Procedure: 1968.

Fig. 19. True call and Flaw detection probabilities for Individual Team A (1968 Ultrasonic Data. corrected 1984-03-07 [31]).

UT Flaw Detection Probability

PVRC-251J (Uncorrected Data)

Team: B. Procedure: 1968.

Fig. 20. True call and Flaw detection probabilities for Individual Team A (1968 Ultrasonic Data. corrected 1984-03-07 [31]).

15

UT Flaw Detection Probability

PVRC-251J (Uncorrected Data)

Team: C.   Procedure: 1968.

Fig. 21. True call and Flaw detection probabilities for Individual Team A (1968 Ultrasonic Data. corrected 1984-03-07 [31]).

UT Flaw-Type Detection Activity Index (DAI) (*)

PVRC-251J (Hedden, 84-03-07)

Team: A.   Procedure: 1968.

(*) No. of true calls divided by no. of possible flaws.
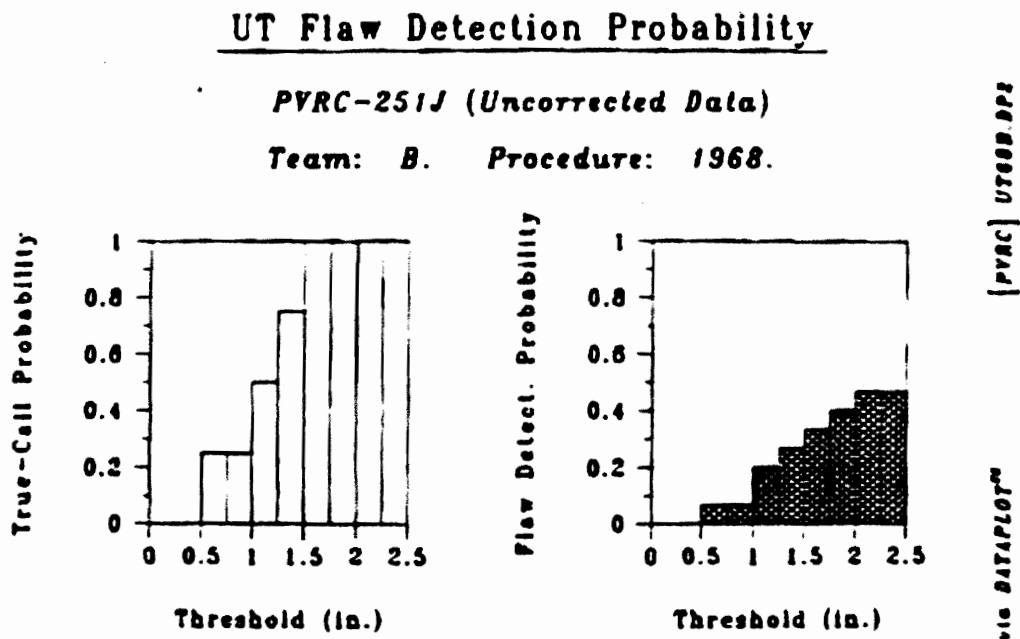
Fig. 22. A plot of Detection Activity Index (DAI) vs. Detection Threshold (DT) as an indicator of **Team A's performance** in detecting 3 different types of flaws using the PVRC 1968 UT Procedure.

UT Flaw-Type Detection Activity Index (DAI) (*)

PVRC-251J (Uncorrected Data)
Team: B. Procedure: 1968.

Cross Cracks   Longitudinal Cracks   Slags

DAI

Threshold (in.)   Threshold (in.)   Threshold (in.)

(*) No. of true calls divided by no. of possible flaws.

Fig. 23. A plot of Detection Activity Index (DAI) vs. Detection Threshold (DT) as an indicator of **Team B's performance** in detecting 3 different types of flaws using the PVRC 1968 UT Procedure.

UT Flaw-Type Detection Activity Index (DAI) (*)

PVRC-251J (Uncorrected Data)
Team: C. Procedure: 1968.

Cross Cracks   Longitudinal Cracks   Slags

DAI

Threshold (in.)   Threshold (in.)   Threshold (in.)

(*) No. of true calls divided by no. of possible flaws.

Fig. 24. A plot of Detection Activity Index (DAI) vs. Detection Threshold (DT) as an indicator of **Team C's performance** in detecting 3 different types of flaws using the PVRC 1968 UT Procedure.

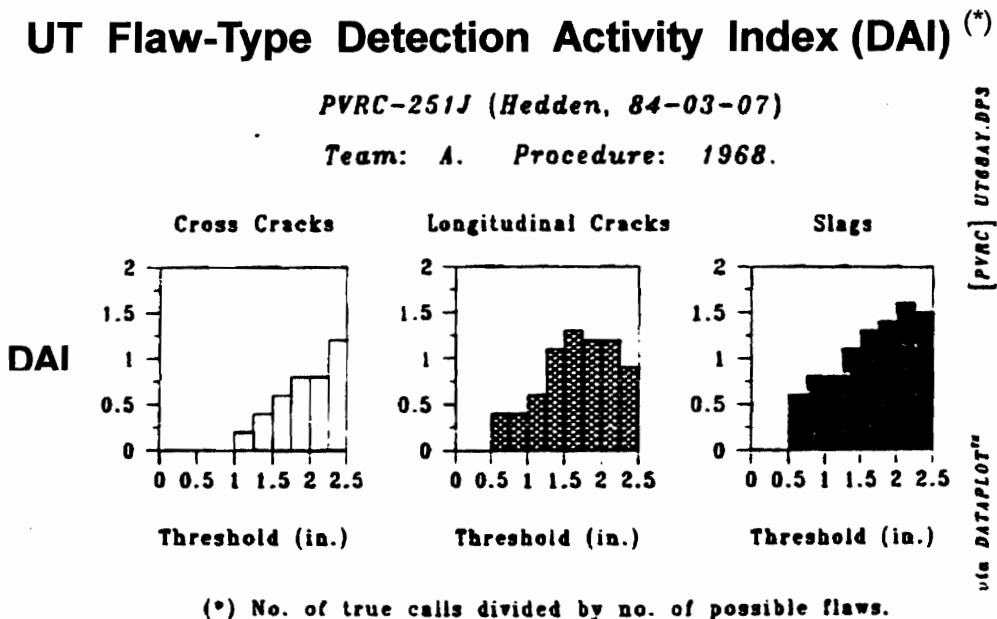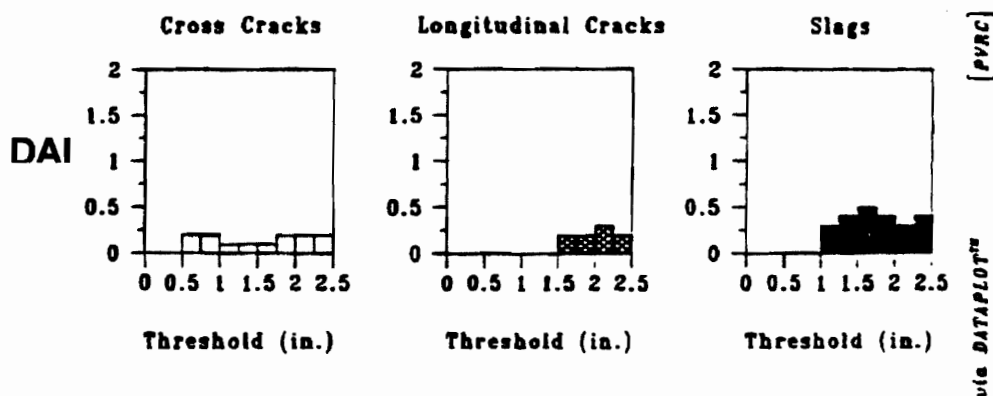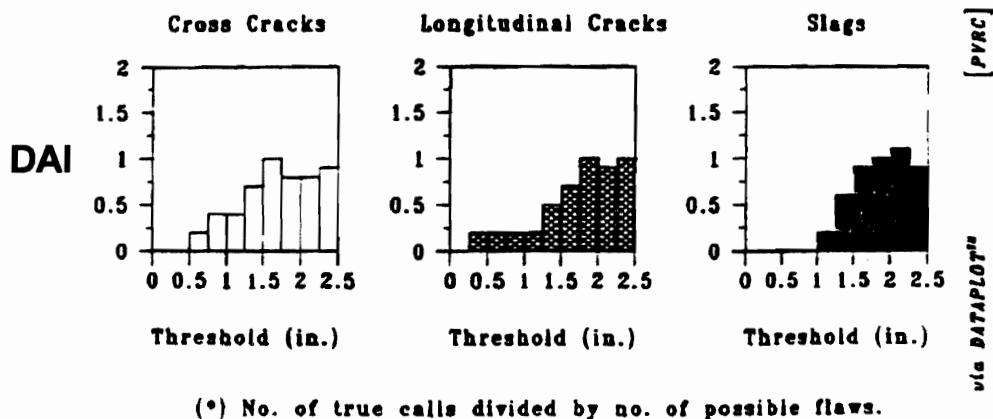## 5. Ultrasonics-1968 Detection Reliability (Continued)

Having defined a detection threshold variable, DT, introduced a mechanism for counting the number of flaws detected for each indication as reported, and developed a call routine of DATAPLOT [25, 26] to automatically analyze and display the detection data such as those given in Figs. 14(a), 15(a), and 17, we began to collect and combine the 36 displays of all three teams to yield a data set of sample size equal to 28, after discarding all indications that failed to detect any beyond a 3-in threshold. A histogram of those 28 data points is given in Fig. 25.

The analysis methodology described in Section 3 was applied to the 28-data set of Fig. 25 to answer the following three questions:

(1)   Is there a between-team effect?

(2)   Is there a flaw-type effect?

(3)   Is there a DAC effect?

Without ANOVA, the estimated sample mean, M, was 1.36 in.,

and the estimated standard deviation, S, was 0.454 in. As shown in Fig. 26, a test for normality of the 28-data set led to the conclusion that the assumption of a normal distribution for the 28-data set was valid. After we conducted a one-way ANOVA for each of the three questions posed earlier, we found practically none in all cases. For example, as displayed in Fig. 25, we answered the first question quantitatively by calculating the standard deviation with team effect to be 0.461 in., which was less than 2% above 0.454 in. For brevity, we omit a similar display for the answer to each of the remaining two questions.

To assess the reliability of the UT-1968 procedure, we followed the same approach as described in Sections 3 and 4 by first obtaining a set of curves for the tolerance factor K for sample size $n = 28$, and then plotting the upper tolerance limits for the 28-data set with $M = 1.36$ in. and $S = 0.454$ in. The result is given in Fig. 27. For a 90% coverage, and a 95% reliability, we found the upper limit of the detection threshold to be 2.35 in. If the acceptance detection threshold criterion of the UT-1968 procedure were set at 2 in., we had to conclude that it failed even though the sample average, 1.36 in., was within 2 in.

UT−1968 Data: Team Effect on Detection

N = 28 .   Ave. = 1.36 .   S.D. = 0.454 .

S.D. (with Replication Effect) = 0.461 .

S.D. (plus Team Effect) = 0.461 .

Number of Flaws Detected

Detection Threshold (in.)
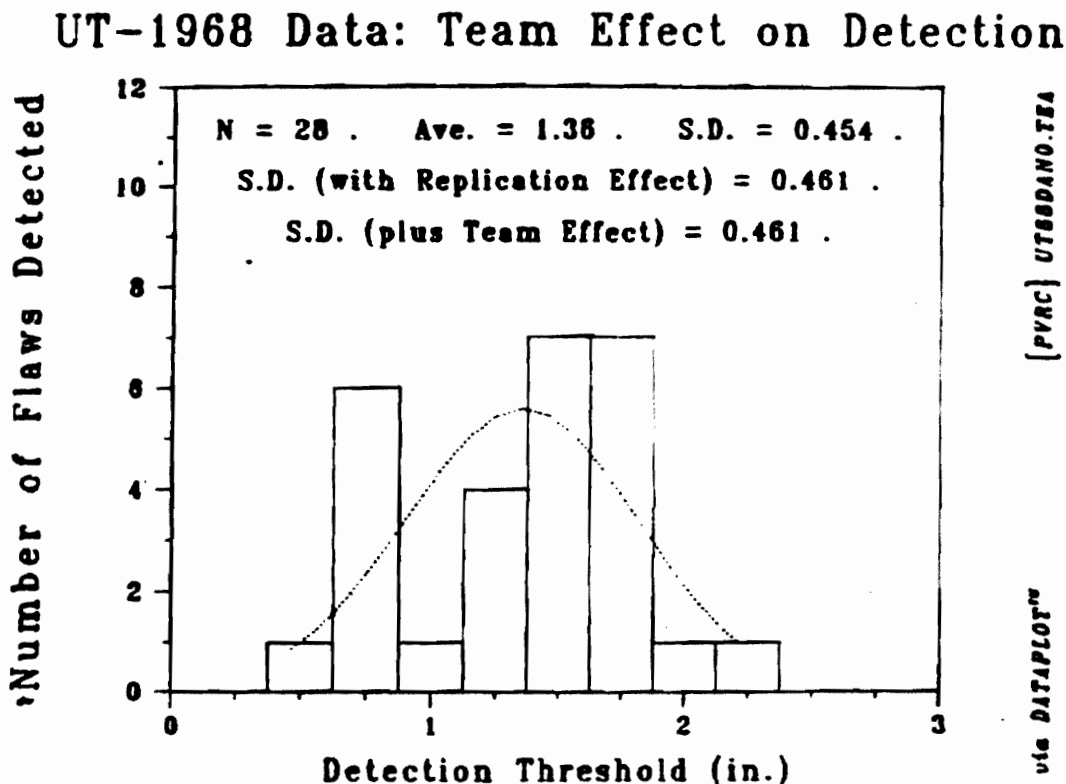
Fig. 25. Results of a one-way analysis of variance (ANOVA) for estimating between-team effect on flaw detection for the 3-team threshold data. Note that no effect was found in the 28-sample data set. Similar results were found for answering the questions whether there were a flaw-type effect and a DAC effect. Plots for those two results are omitted for brevity.

18

Test for Normality
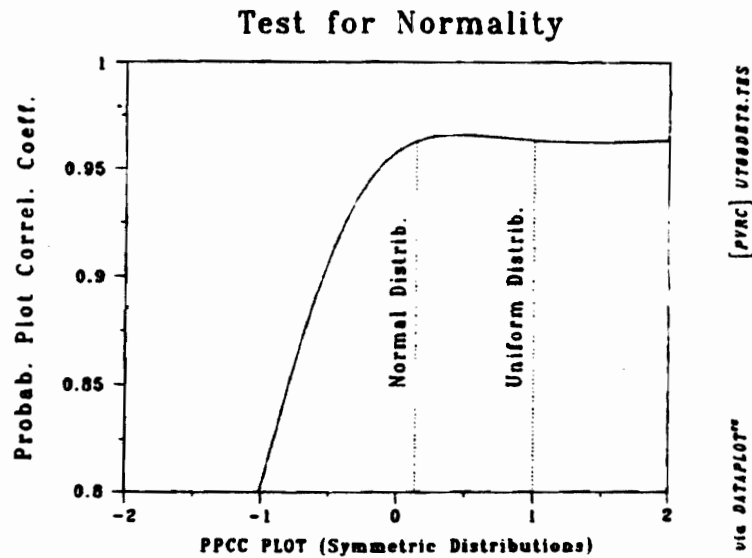
Fig. 26. Lambda Test (Filliben [34, 35]) for Symmetric Distribution Fit of the 3-team UT-1968 Detection Threshold Data.

UT-1968 Detection Reliability

No. of Flaws = 15. No. of Types = 4.
No. of Teams = 3. No. of Hits = 28.
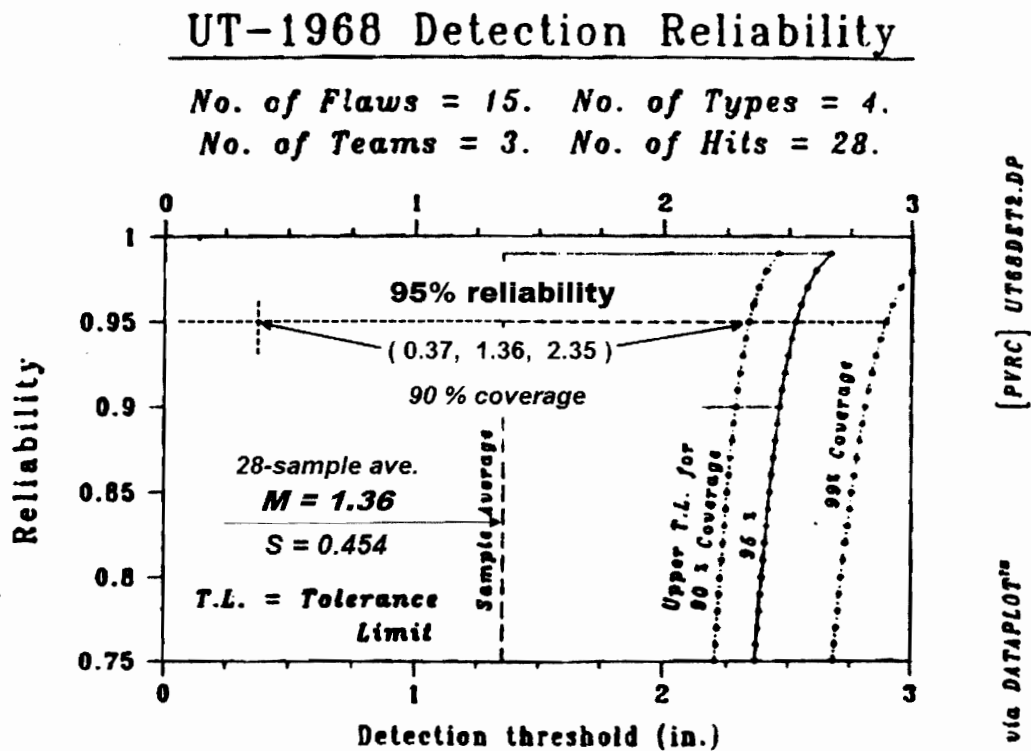
Fig. 27. A plot of the Reliability of Ultrasonic Detection (PVRC 1968 Procedure) of Weld Flaws as defined by the upper tolerance limit curves vs. Detection Threshold (DT) for three different coverages, $P = 90\%$, 95%, and 99%.

19

6. DESIGN OF EXPERIMENTS APPROACH TO NDE UNCERTAINTY ESTIMATION

The analysis methodology described in Section 3, as originally reported in Refs. [12-14] and applied to the assessment of reliability of two PVRC procedures as reviewed in Sections 4 and 5, depended on two classical statistical data analysis techniques, namely, the ANOVA and the tolerance limit intervals. The methodology is capable of delivering an answer to a user on the question whether a specific procedure is acceptable based on a given criterion and the results of a round-robin test program. But the methodology does not go far enough to prescribe an efficient and rigorous follow-up plan of how to do more tests, if and when the procedure fails.

In this section, we introduce a third statistical data analysis technique, known as the design of experiments (DOE) as described in Box, Hunter, and Hunter [46], and Montgomery [47], that will rectify the shortcomings of the previous methodology. For a detailed exposition of this DOE technique, the reader may consult Croarkin, et al [48], or a recent tutorial paper with two examples by Fong, et al [49].

To acquaint the reader with the fundamentals of DOE, we present below a series of 8 questions and answers as a background for understanding an NDE example of DOE. (Note: For a reader familiar with the DOE technique, one can skip this part and go directly to the example that follows.]:

DOE 1.1   What is design of experiments (DOE)?
Ans.   In an experiment, we change one or more process variables (factors) in order to observe the effect the changes have on one or more response variables. DOE is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions.

DOE begins with determining the *objectives* of an experiment and selecting the *process factors* for the study. An Experimental Design is the laying out of a detailed experimental plan in advance of doing the experiment. Well chosen experimental designs maximize the amount of "information" that can be obtained for a given amount of experimental effort.

DOE 1.2   What is the first step in applying the DOE method?
Ans.   The statistical theory underlying DOE begins with the concept of *process models*. A process *model* of the 'black box' type is formulated with several discrete or continuous input *factors* that can be controlled, and one or more measured output *responses*. The output responses are assumed continuous. Real or virtual experimental data are used to derive an empirical (approximate) model linking the outputs and inputs. These empirical models generally contain *first-order (linear) and second-order (quadratic) terms.*

DOE 1.3   What is a first order model?
Ans.   A first-order model with three factors, X_1, X_2 and X_3, can be written as

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_{12}X_1X_2+\beta_{13}X_1X_3+\beta_{23}X_2X_3+errors \quad (4)$$

Here, Y is the response for given levels of the main effects X_1, X_2 and X_3, and the X_1X_2 , X_1X_3 , X_2X_3 terms are included to account for a possible interaction effect between X_1 and X_2, X_1 and X_3 , X_2 and X_3 , respectively. The constant β_0 is the response of Y when both main effects are 0. In the example that follows for an application in NDE, we use a linear model with five factors and one response variable, and the total number of terms on the right hand side of eq. (4) is 2^5, or 32.

DOE 1.4   How does one select factors and responses?
Ans.   Process variables of an experiment include both inputs (factors) and outputs (responses). The selection criteria are:

(a) Include all important factors (based on judgment).
(b) Be bold in choosing the low and high factor levels.
(c) Check the factor settings for impractical or impossible combinations, such as very low pressure or very high gas flows.
(d) Include all relevant responses.
(e) Avoid using only responses that combine two or more measurements of the process. For example, if interested in the ratio of two rates, measure both rates, not just the ratio.

We have to choose the range of the settings for input factors, and it is wise to give this some thought beforehand rather than just try extreme values.

DOE 1.5   How does one select an experimental design?
Ans.   The most popular experimental designs are *two-level designs*. Why only two levels? There are a number of good reasons why two is the most common choice amongst engineers; one reason is that it is ideal for screening designs, simple and economical; it also gives most of the information required to go to a multilevel response surface experiment if one is needed.

The standard layout for a 2-level design uses +1 and -1 notation to denote the "high level" and the "low level" respectively, for each factor. For example, the matrix below

	Factor 1 (X1)	Factor 2 (X2)
Trial 1	-1	-1
Trial 2	+1	-1
Trial 3	-1	+1
Trial 4	+1	+1

describes an experiment in which 4 trials (or runs) were conducted with each factor set to high or low during a run according to whether the matrix had a +1 or -1 set for the factor during that trial. If the experiment had more than 2 factors, there would be an additional column in the matrix for each additional factor.

6. *Design of Experiments Approach (Continued)*

DOE 1.6   What is a 2-level full factorial DOE?
Ans.   A common experimental design is one with all input factors set at two levels each. These levels are called 'high' and 'low', or '+1' and '-1', respectively. A design with all possible high/low combinations of all the input factors is called a full factorial design of experiments in two levels.

If there are k factors, each at 2 levels, a full factorial DOE has 2^k runs. Fig. 28 is a graphical representation of a 2-level, 3-factor, 2^3 or 8-run full factorial DOE. This implies eight runs (not counting replications or center point runs). The arrows show the direction of increase of the factors. The numbers '1' through '8' at the corners of the design box reference the "Standard Order" of runs (also referred to as the "Yates Order", see Ref. [48]). When the number of factors is 5 or greater, a full factorial DOE requires a large number of runs and is not very efficient. This is where a need for a fractional factorial DOE comes in.

DOE 1.7   What is a Center Point in a 2-level design?
Ans.   To introduce the concept of a center point, we again refer to Fig. 28, a graphical representation of a two-level, full factorial design for three factors, namely, the 2^3 design.

As mentioned earlier, we adopt the convention of +1 and -1 for the factor settings of a two-level design. When we include a center point during the experiment, we mean a point located in the middle of the design cube, and the convention is to denote a center point by the value "0".

DOE 1.8   What is a 2-level fractional factorial DOE?
Ans.   A fractional factorial DOE is a factorial experiment in which only an adequately chosen fraction of the treatment combinations required for the complete factorial experiment is selected to be run. In general, we pick a fraction such as ½, ¼, etc. of the runs called for by the full factorial. We use various strategies that ensure an appropriate choice of runs. Properly chosen fractional factorial designs for 2-level experiments have
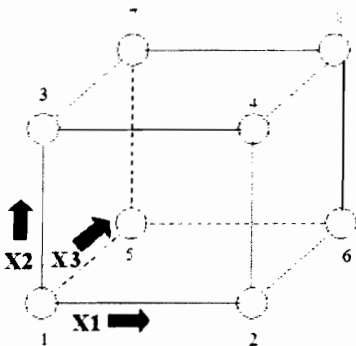

Fig. 28. A 2^3 2-level, full factorial design; factors X1, X2, X3.

the desirable properties of being both *balanced* and *orthogonal*.

A NDE Example of DOE
Let us consider a. fictitious scenario where an ultrasonic testing team such as Team A of the PVRC 1968 UT round robin program reported a single indication, namely, #1, as shown in Figs. 14(a) and 14(b), where they found a crack of 2.0 in. long, whereas the actual crack, "L", was 4.75 in. long and located at 0.65 in. away, center-to-center. The question is:

What additional measurements should Team A make in order to report the result, 2.0 (?) in, with a 2-sided predicted 95% level of confidence?

For simplicity, we do not ask the deeper question of a 95% reliability with a coverage of P, because that will involve an extra step of computing the tolerance factors as we did in Sections 3, 4, and 5, a distraction in our goal to introduce DOE.

To answer this question, we need to select a response variable (output Y), several factors (input $X1$, $X2$, ...) of importance in ultrasonic testing, the high and low settings of each factor, and a 2-level fractional factorial design. Clearly, the detected crack length is an appropriate response variable, so we denote the length by $Y1$. For this exercise, we choose the following factors as appropriate for a UT follow-up experiment:

Factor	Title (Unit)	Low	Center	High
X1	Operator's Experience (Year)	2.0	4.0	6.0
X2	UT Machine Age (Year)	2.0	5.0	8.0
X3	Cable Length (feet)	6.0	8.0	10.0
X4	Transducer Probe Angle (deg.)	42.0	45.0	48.0
X5	Plastic Shoe Thickness (in.)	0.25	0.50	0.75

We also choose a 5-factor 8-run+center-point fractional factorial design as shown in Fig. 29, where the center point response variable $Y1$ equals 2.0 in., as reported by Team A.

Without actually obtaining the 8 additional measurements as proposed, we now conduct a "virtual" experiment by arbitrarily assign 8 values of $Y1$ to those 8 runs as shown in Fig. 29. Using a 10-step exploratory-data-analysis routine of DATAPLOT [26], we obtain results in the form of plots and tables as shown in Figs. 29 through 35.

In particular, Fig. 29 is a display of the 8 values of $Y1$ as an ordered set. Fig. 30 is a main effects plot where X1 and X4 are shown to be dominant. Fig. 31 is an interaction effects plot where 7 of the 10 two-term interactions, X1X2, X2X3, X2X4, X2X5, X3X4. X3X5. and X4X5, appear to be confounding. In Fig. 32, we note that only X2X3 and X2X5 are quantitatively significant, and since X2 as a main effect is zero, we conclude that X2X3 and X2X5 are not strong enough to interfere with our decision to choose X1 and X4 as dominant. Fig. 33 is a 2-parameter least square fit, and Fig. 34 is an uncertainty analysis. Fig. 35 is a contour plot of the 2 dominant factors with the important answer to our question, i.e., the fictitious DOE yields a predicted crack length equal to 2.05 (0.54), or 2.05 ± 0.54 in., with 95% confidence.

21

Virtual 9-run UT Experiment for Detecting Flaw "L"

Step 1

k = 5
n = 8

Y1	X1	X2	X3	X4	X5
1.70	- 1	- 1	- 1	+1	+1
2.15	+1	- 1	- 1	- 1	- 1
1.75	- 1	+1	- 1	- 1	+1
2.65	+1	+1	- 1	+1	- 1
2.25	- 1	- 1	+1	+1	- 1
2.10	+1	- 1	+1	- 1	+1
1.30	- 1	+1	+1	- 1	- 1
2.50	+1	+1	+1	+1	+1

Response Y (Length of Crack L (in))

Ordered Data Set

X1:	·	·	·	+	+	·	+	+	Serv Year SY
X2:	+	·	+	·	·	·	+	+	Machine Y MY
X3:	+	·	·	+	·	+	+	·	Cable Lgh CL
X4:	·	+	·	·	·	+	+	+	Prb Angle PA
X5:	·	+	+	+	·	·	+	·	Shoe Thck ST

Factors: X1 = Service Yr (2, 4, 6); X2 = Machine Yr (2, 5, 8); X3 = Cable Length (6, 8, 10); X4 = Probe Angle (42, 45, 48); X5 = Shoe Thickness (1/4", 1/2", 3/4"); Values (-, 0, +)
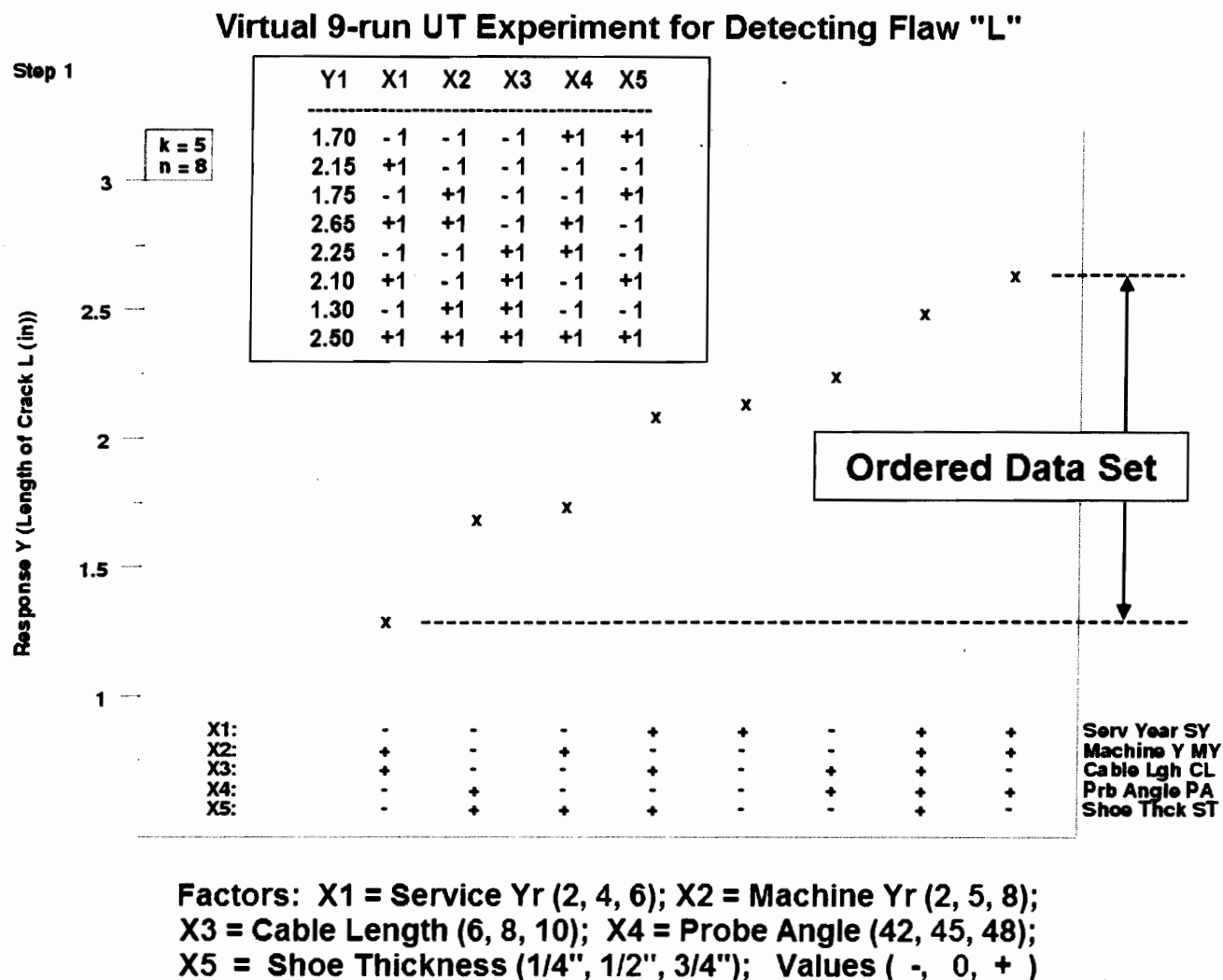
Fig. 29. First of ten plots by DATAPLOT showing the fictitious ultrasonic crack measurement data as an ordered set. Note the table at the bottom of the plot being the transposed DOE matrix with re-ordered columns.

Virtual 9-run UT Experiment for Detecting Flaw "L"

Main Effects Plot

Step 3

Factors: X1 = Service Year (2, 4, 6); X2 = Machine Year (2, 5, 8); X3 = Cable Length (6, 8, 10); X4 = Transducer Probe Angle (42, 45, 48); X5 = Plastic Shoe Thickness (1/4", 1/2", 3/4"); Values (-, 0, +)
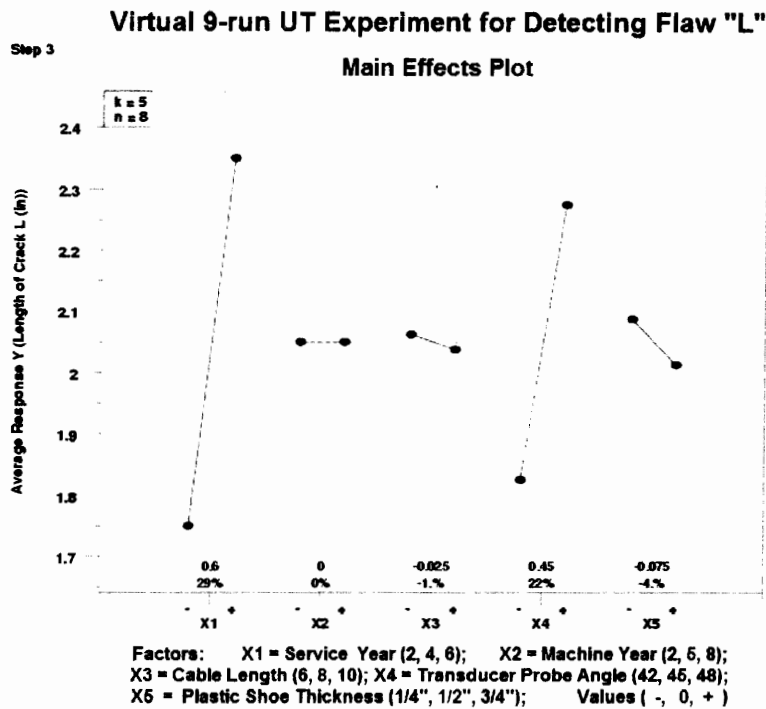
Fig. 30. Step 3 of a 10-step analysis of the fictitious ultrasonic crack measurement data showing the main effects of the 5-factor 8-run fractional factorial 2-level design (k = 5, n = 8). Note X1 and X4 are dominant.

Virtual 9-run UT Experiment for Detecting Flaw "L"

Interaction Effects Matrix
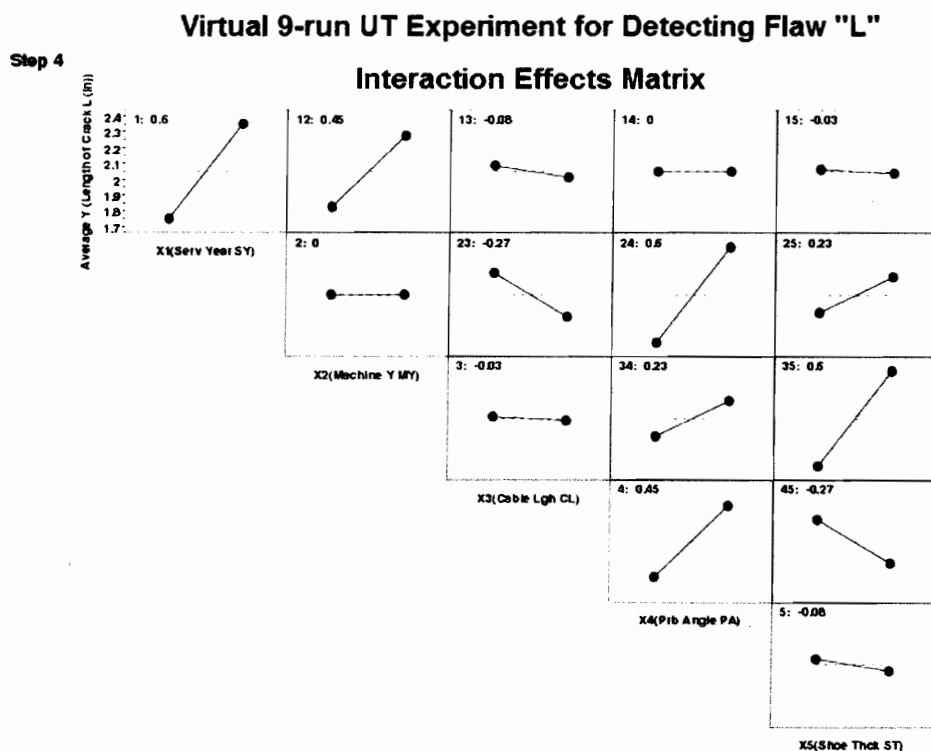
Step 4

Fig. 31. Step 4 of a 10-step analysis of the fictitious ultrasonic crack measurement data showing the interaction effects of the 5-factor 8-run fractional factorial 2-level design (k = 5, n = 8). Note the existence of 7 out of 10 potential 2-term interactions that need to be addressed in Step 7, Fig. 32.

6. Design of Experiments Approach (Continued)

Step 7

Virtual 9-run UT Experiment for Detecting Flaw "L"

|Effects| Plot

Average = 2.05

Factor: Confound

1: 1+24+35

4: 4+12

Factor	Effect
1+24+35	: 0.6
4+12	: 0.45
23+45	: -0.275
25+34	: 0.225
5+13	: -0.075
3+15	: -0.025
2+14	: 0

|Effects|

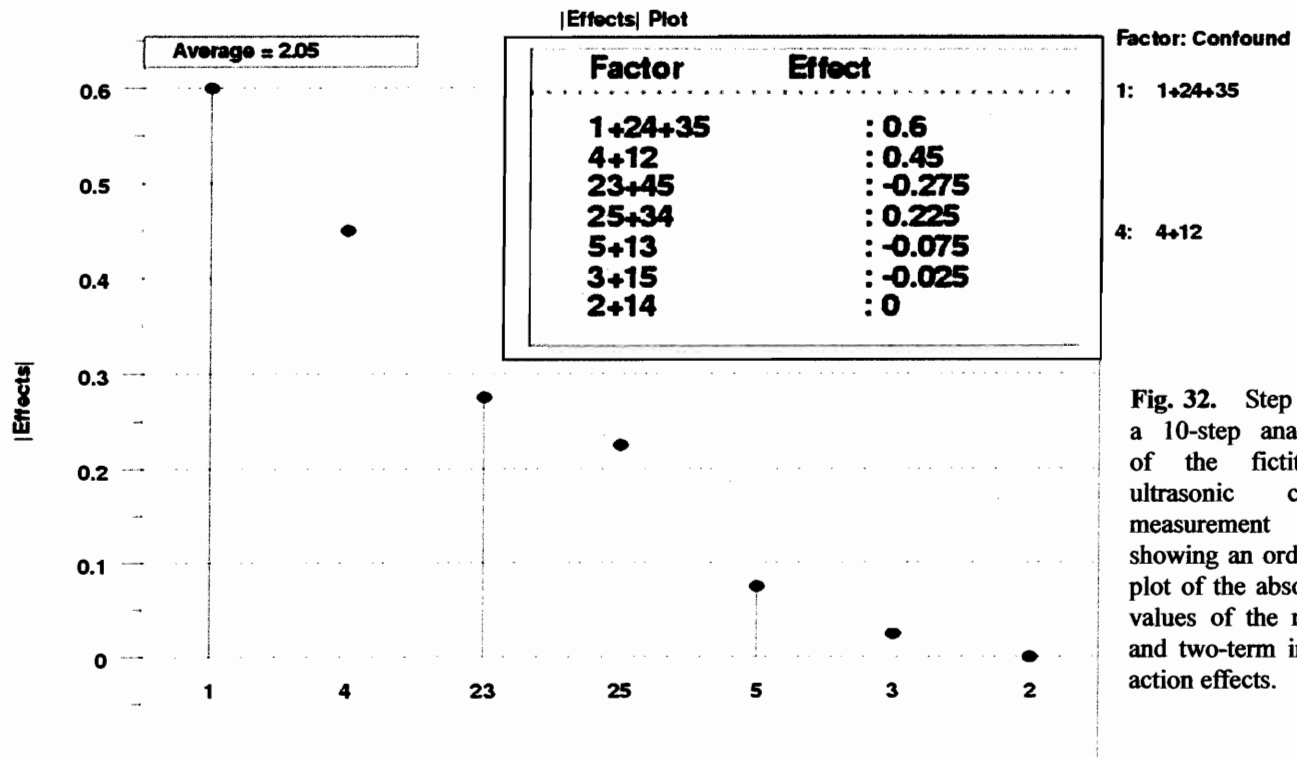0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0

1   4   23   25   5   3   2

Fig. 32. Step 7 of a 10-step analysis of the fictitious ultrasonic crack measurement data showing an ordered plot of the absolute values of the main and two-term interaction effects.

Factor 1 = Serv Year, 2 = Machine Y, 3 = Cable Length, 4 = Probe Angle, 5 = Shoe Thickness; 23 = interaction,

```
                    LEAST  SQUARES  MULTILINEAR  FIT

SAMPLE SIZE N      =    9            NUMBER OF VARIABLES      =     2

REPLICATION CASE:       REPLICATION STANDARD DEVIATION   =    0.2573908
                        REPLICATION DEGREES OF FREEDOM    =    4
                        NUMBER OF DISTINCT SUBSETS        =    5

            PARAMETER ESTIMATES          (APPROX. ST. DEV.)      T VALUE
    1   A0                  2.04444       (0.7035E-01)           29.06
    2   A1        X1        0.300000      (0.7461E-01)            4.021
    3   A2        X4        0.225000      (0.7461E-01)            3.016

RESIDUAL     STANDARD DEVIATION =            0.2110380
RESIDUAL     DEGREES OF FREEDOM =            6

REPLICATION STANDARD DEVIATION =            0.2573907673
REPLICATION DEGREES OF FREEDOM =            4

LACK OF FIT F RATIO   =    0.0168    = THE    1.6563%   POINT OF THE
                     F DISTRIBUTION WITH 2 AND 4 DEGREES OF FREEDOM
```
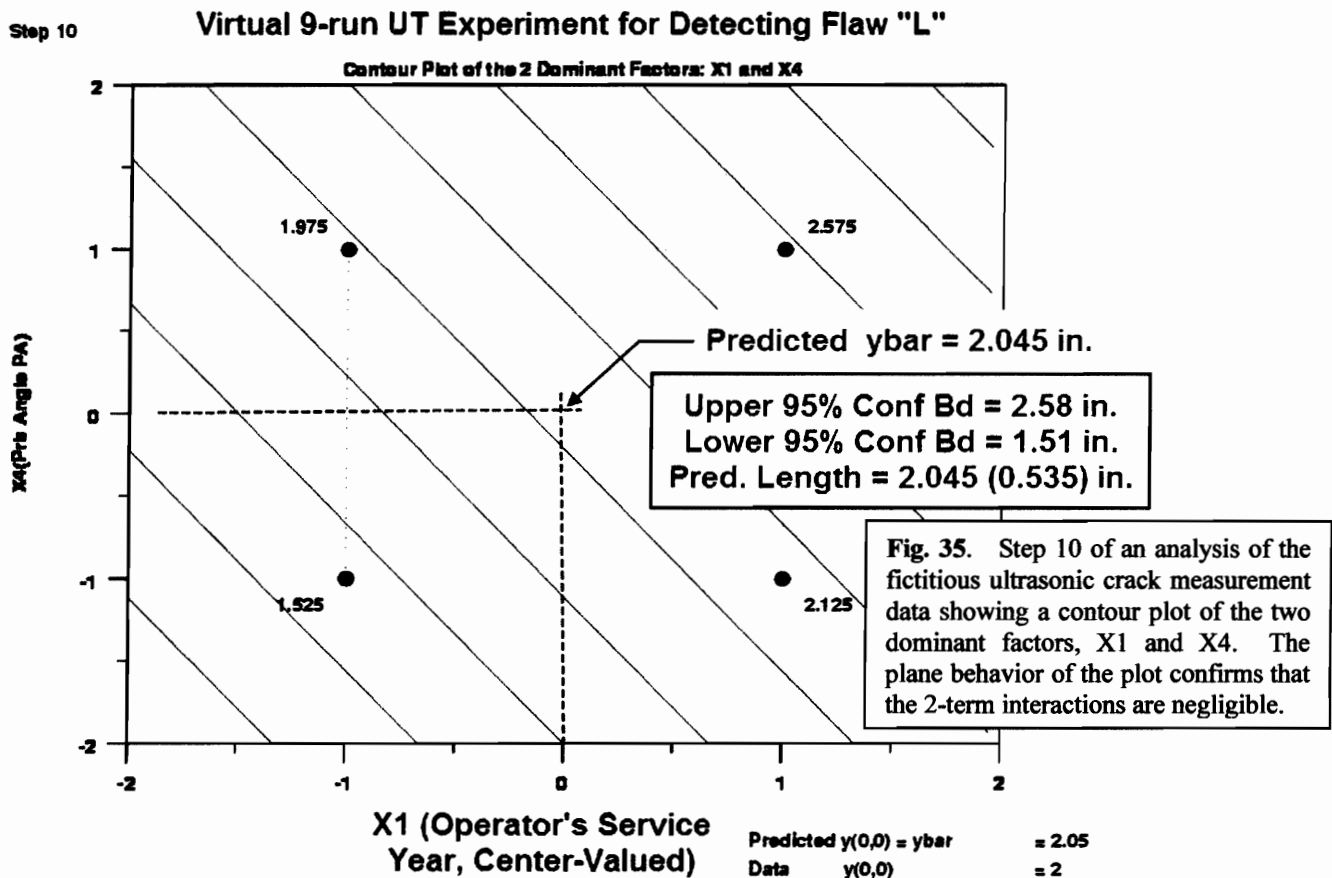
Fig. 33. Sensitivity analysis of the 5-factor 8-run-plus-center-point fictitious ultrasonic crack measurement data using a two-parameter (X1, X4) least square fit routine of DATAPLOT.

24

```
                  Uncertainty Analysis

A0                                  =  2.044445
A1                                  =  0.3
A2                                  =  0.225

Resdiual  SD                        =  0.211038
Resdiual  DF                        =  6

Variance(Y)                         =  0.04258
SD(Y)                               =  0.206347

Upper  95%  Confidence  Bound  for  Y  =  2.57667
Lower  95%  Confidence  Bound  for  Y  =  1.51222
```

Fig. 34. Results of an uncertainty analysis generated by DATAPLOT 10-step code showing that the UT Team A's detection of a crack has a reportable length within a 95% confidence interval of (1.51 in., 2.58 in.).

Step 10

Virtual 9-run UT Experiment for Detecting Flaw "L"

Contour Plot of the 2 Dominant Factors: X1 and X4

Predicted ybar = 2.045 in.

Upper 95% Conf Bd = 2.58 in.
Lower 95% Conf Bd = 1.51 in.
Pred. Length = 2.045 (0.535) in.

X4(Prb Angle PA)

X1 (Operator's Service Year, Center-Valued)

Predicted y(0,0) = ybar = 2.05
Data   y(0,0)   = 2

Fig. 35. Step 10 of an analysis of the fictitious ultrasonic crack measurement data showing a contour plot of the two dominant factors, X1 and X4. The plane behavior of the plot confirms that the 2-term interactions are negligible.

25

7. A PROTOTYPE WEB-BASED DATA ANALYSIS PLUG-IN FOR REMOTE-ACCESS OF DATAPLOT

Since NDE monitoring of aging structures is usually performed in the field and the analysis of the field-generated NDE data is accomplished in the office, there is a time delay in getting results of the analysis back to the field for maintenance decision making. To reduce this time delay, a demonstration project was initiated at the National Institute of Standards and Technology (NIST), where a prototype web-based data analysis "plug-in" was implemented using DATAPLOT as the software package for remote-access testing. In this paper, we shall describe how this plug-in works in a step-by-step presentation where DATAPLOT is accessed remotely and executed to do a similar task as the ultrasonic detection data analysis of Sect. 6.

In Fig. 36, we show the results of three steps, namely, Step 1 begins with the url, www.nist.gov and a change of the url to the personal website of Alan Heckert, Step 2 is to activate the link named "e-Metrology," and Step 3 is to activate the link named "Experimental Design." In Fig. 37, we show Step 4 where we click on the link named "10 Step Analysis of Factorial Designs with Uncertainty." In Figs. 38, 39, and 40, we show Step 5 in three parts, because the screen contains detailed instructions on how to do the analysis and a template for filling 2 major and 8 minor dialog boxes with input data.

In Fig. 41, we show Step 6, which is not a web-based task, but is important in getting ready for Step 7 by preparing an ASCII file containing all the necessary input data for the 10-step analysis. In Figs. 42 and 43, we show Step 7 in two parts, where the data of Step 6 is now entered. After clicking on "Perform 10-Step Analysis" at the end of Step 7, we reach the final Step 8 as shown in Fig. 44 where the first plot of the results of the 10-step analysis appears. This plot is similar to Fig. 29 of Sect. 6. The remaining plots and tables of the 10-step analysis results then follow. Task accomplished.
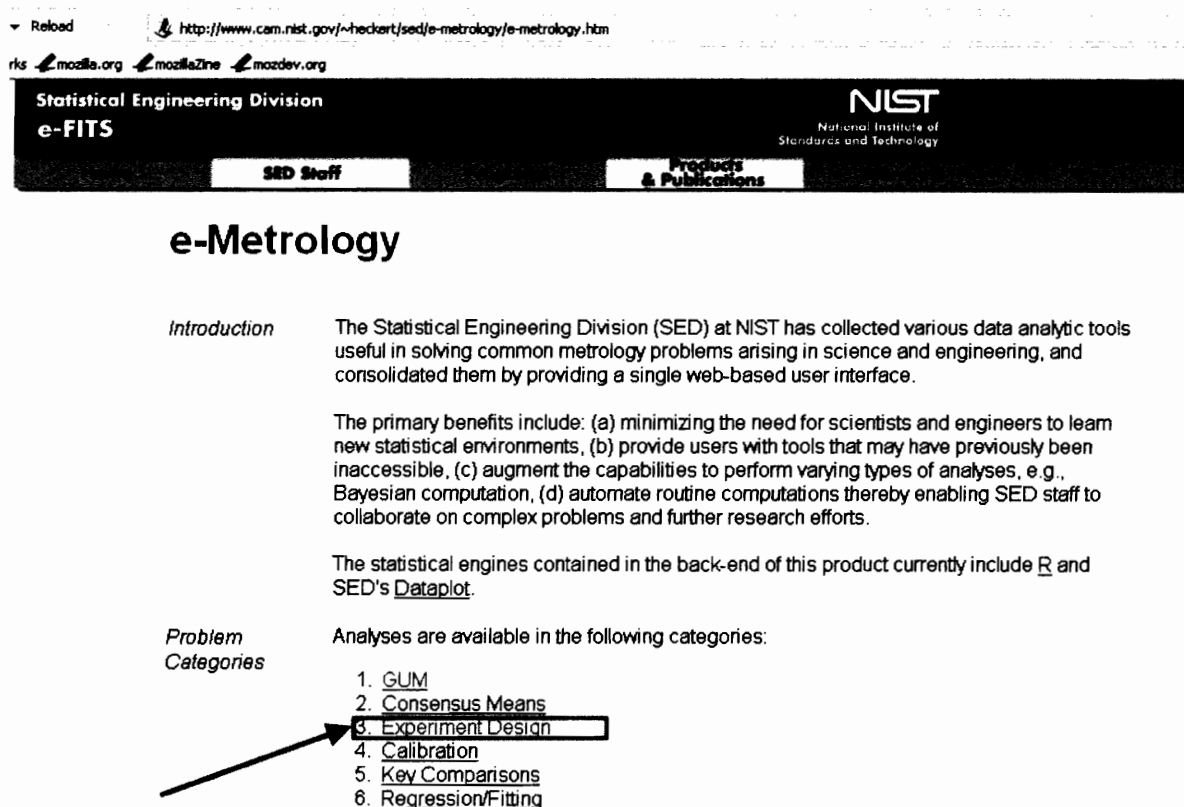

Fig. 36. Step 3 of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

http://www.cam.nist.gov/~heckert/sed/e-metrology/e-metrology.htm#dex

'.mozillaZine mozdev.org

• **Design of Experiments**

Analytic	Description
10-Step Analysis of Factorial Designs	Perform the 10-step exploratory data analysis of 2-level full factorial and fractional factorial designs.
10-Step Analysis of Factorial Designs with Uncertainty	Perform the 10-step exploratory data analysis of 2-level full factorial and fractional factorial designs. In addition, include some uncertainty computations.

Fig. 37. Step 4 of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

Reload   http://www.cam.nist.gov/~heckert/sed/e-metrology/10stepunc.htm

rks mozilla.org mozillaZine mozdev.org

Statistical Engineering Division
Web Based Computing

NIST
National Institute of
Standards and Technology

SED Staff

Products
& Publications
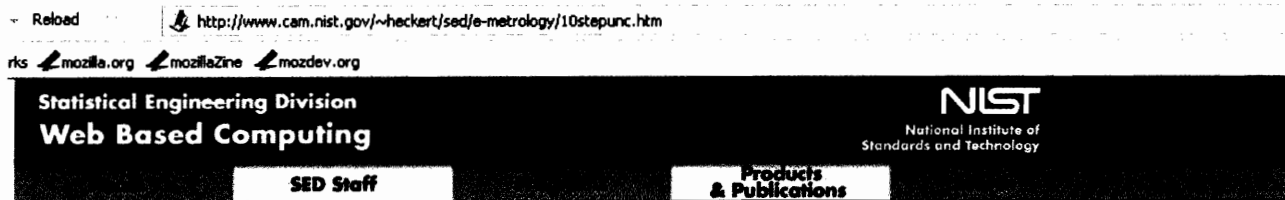
e-Metrology: 10-Step EDA Analysis of 2-Level Full and Fractional Factorial Designs

Perform a 10-Step Analysis of 2-Level Full and Fractional Factorial Designs

Note: This is a big screen with notes and dialog boxes.

Overview	An important class of experiment designs are the full factorial and the fractional factorial designs.
Full Factorial Designs	A design with all possible combinations of all the input factors is called a full factorial design.
Fractional Factorial Designs	If only a subset of the full factorial design points are used, then we have a fractional factorial design.
	Fractional factorial designs allow us to run experiments with a larger number of factors.

We will download in parts. This is

Part 1

Fig. 38. Step 5, Part 1, of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

27

http://www.cam.nist.gov/~heckert/sed/e-metrology/10stepunc.htm

mozillaZine mozdev.org

Fractional Factorial Designs	If only a subset of the full factorial design points are used, then we have a fractional factorial design. Fractional factorial designs allow us to run experiments with a larger number of factors.
2-Level Designs, Center Points	An important subclass of these designs are the 2-level designs.
10-Step EDA Approach	This form analyzes 2-level full and fractional factorial designs using a 10-step EDA approach developed by Jim Filliben.
Yates Order	This analysis assumes that your response variable is in Yates order.

Note: This is a big screen with notes and dialog boxes.

We will download in parts. This is Part 2

Note the two dialog boxes below:

Enter the response data (Yates order, no center points):

Enter the design matrix (Yates order, no center points):

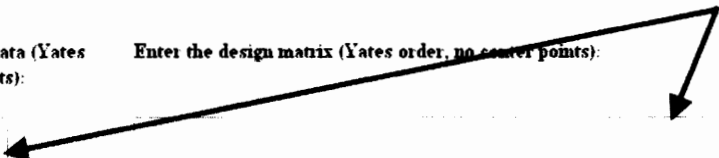**Fig. 39.** Step 5, Part 2, of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

http://www.cam.nist.gov/~heckert/sed/e-metrology/10stepunc.htm

mozillaZine mozdev.org

Enter the response data (Yates order, no center points):

Enter the design matrix (Yates order, no center points):

Note: This is a big screen with notes and dialog boxes.

We will download in parts. This is Part 3

We need to fill the two dialog boxes with data (see next slide for example)

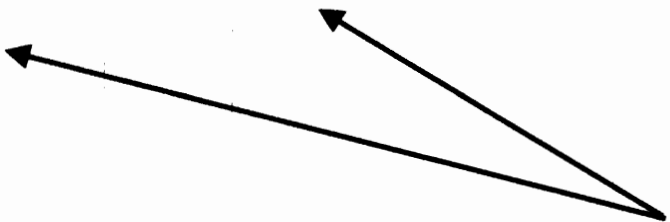**Fig. 40** . Step 5, Part 3, of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

```
END OF DATA

THE ABOVE IS AN ASCII DATA FILE TO BE READ AS INPUT TO A DATAPLOT FILE.

ALL LINES AFTER THE COMMAND, "END OF DATA", WILL BE IGNORED.|
-----------------------------------------------------------------

   ASCII LINES AS INPUT DATA FOR WEB-BASED DATAPLOT APPLICATION:

  http://www.cam.nist.gov/~heckert/sed/e-metrology/e-metrology.htm

    AN EXAMPLE FOR FILLING  2  MAJOR AND  8  MINOR DIALOG BOXES
-----------------------------------------------------------------

MAJOR BOX 1.  RESPONSE VARIABLE WITHOUT CENTER POINT

    111.310
     89.100
     96.020
    112.520
    144.500
    132.520
    128.280
    156.400


MAJOR BOX 2.  DESIGN MATRIX

    -1.  -1.  -1.   1.       1.
     1.  -1.  -1.  -1.      -1.
    -1.   1.  -1.  -1.       1.
     1.   1.  -1.   1.      -1.
    -1.  -1.   1.   1.      -1.
     1.  -1.   1.  -1.       1.
    -1.   1.   1.  -1.      -1.
     1.   1.   1.   1.       1.
```

---▶ Note: This is an ASCII file containing data as input to a DATAPLOT code for a 10-step analysis of factorial design with uncertainty.

After we download the web screen Part 3, we need this ASCII file to fill 2 major and 8 minor dialog boxes

```
MINOR BOX 1.  center point value  =    120.840

MINOR BOX 2.  title = Sensitivity Analysis on Chord 708 section force (kipf)

MINOR BOX 3.  RESPONSE LABEL      =   Chord 708 section force (kipf)

MINOR BOX 4.  Factor 1            =   Elas Mod  E

MINOR BOX 5.  Factor 2            =   Thickness  b

MINOR BOX 6.  Factor 3            =   Nodal Load P

MINOR BOX 7.  Factor 4            =   Settlemt U2C

MINOR BOX 8.  Factor 5            =   Settlemt U1C
```

Fig. 41. Step 6 of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

111.310		-1.	-1.	-1.	1.	1.
89.100		1.	-1.	-1.	-1.	-1.
96.020		-1.	1.	-1.	-1.	1.
112.520		1.	1.	-1.	1.	-1.
144.500		-1.	-1.	1.	1.	-1.
132.520		1.	-1.	1.	-1.	1.
128.280		-1.	1.	1.	-1.	-1.
156.400		1.	1.	1.	1.	1.

Enter center point values (if any):   120.840

Enter the title across the top of each page: ysis on Chord 708 section force (ki

Enter the sub-title across the top of each page:

Enter the label for the response variable   Chord 708 section force (kipf)

Enter the label for factor variable 1:   Elastic Modulus E

Enter the label for factor variable 2:   Thickness b

Enter the label for factor variable 3:   Nodal Load P

Enter the label for factor variable 4:   Settlement U2C

Enter the label for factor variable 5:   Settlement U1C

Fig. 42. Step 7, Part 1, of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

Enter the label for factor variable 1:   Elastic Modulus E

Enter the label for factor variable 2:   Thickness b

Enter the label for factor variable 3:   Nodal Load P

Enter the label for factor variable 4:   Settlement U2C

Enter the label for factor variable 5:   Settlement U1C

Enter the label for factor variable 6:

Enter the label for factor variable 7:

Enter the label for factor variable 8:

Enter the label for factor variable 9:

Enter the label for factor variable 10:

Perform 10-Step Analysis   Reset

Fig. 43. Step 7, Part 2, of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

```
At the Beginning of DEXPLOTSUB.DP . . .


At the Beginning of CONFOUND.DP . . .

k,n = 5,8


At the Beginning of Step  1 (DEXODP.DP) . . .
```

Note: This is the first of 12 pages of a .pdf file containing results of the 10-step analysis [1]
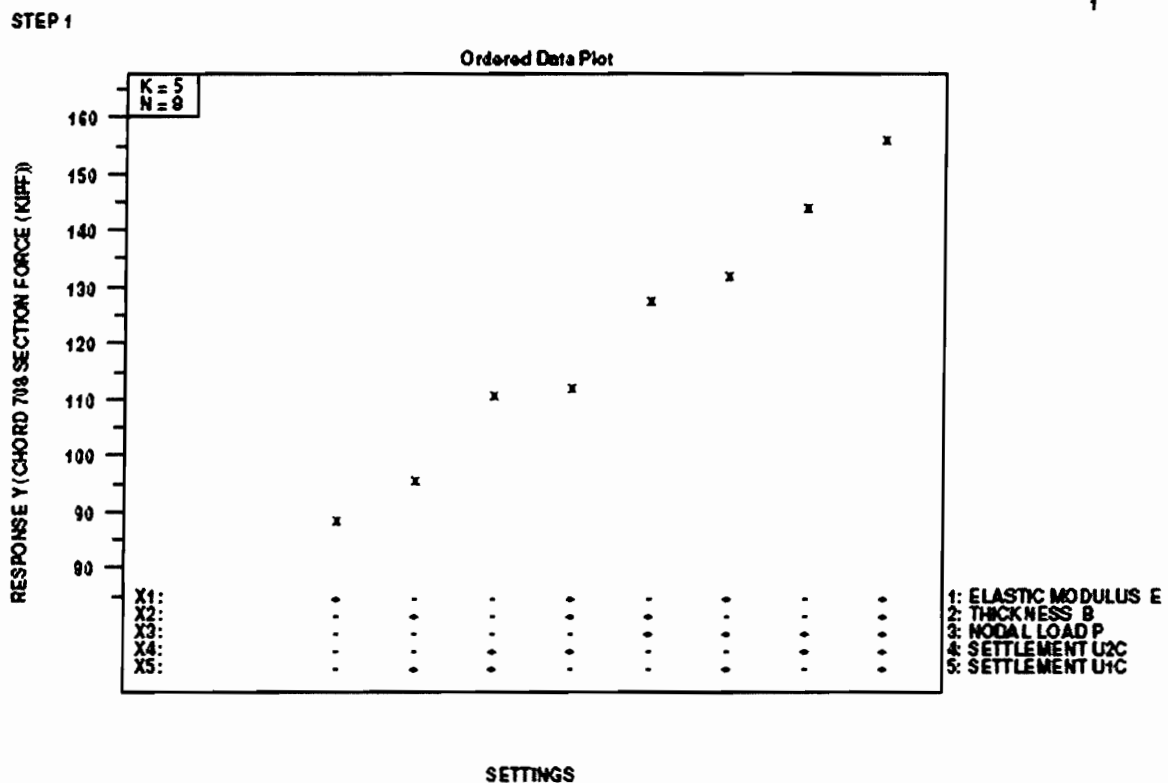
STEP 1

Fig. 44.   Step 8 of a Prototype Web-based Data Analysis Plug-in for Estimating Reliability of Network-Accessed Testing Data

8. SIGNIFICANCE AND LIMITATIONS

This paper introduces two computer-based tools to the engineering community, not only for estimating reliability of weld flaw detection, location, and sizing as stated in the title of the paper, but also for a more general audience, where one can use the DOE tool to convert experience and judgment into uncertainty estimation for better decision making, and the web-based tool to accelerate critical analysis by getting input from the field, computing in the office and relaying predictions and recommendations back to the field within hours instead of days.

By using NDE applications as examples to illustrate two statistical data analysis methods, one on ANOVA and tolerance intervals, and the other on DOE, we believe this paper is a significant contribution to the advancement of NDE engineering and science for preventing premature failures of aging structures and components. In particular, the NDE example of DOE as presented in Section 6 included an interesting result by identifying the UT operator's experience as one of the two dominant factors. This coincided fortuitously with a similar observation by Behravesh and Dau [50] in a 1986 EPRI report on NDE personnel requalification, even though our example used fictitious data that had no physical basis.

Clearly, the two statistical data analysis methods as applied to NDE are not without limitations. Chief among them is the difficulty of reducing a large list of NDE-related factors to a small and manageable number. In addition, the search for one or more critical response variables to characterize an NDE process is also a challenging one. The third limitation is the judgment-based requirement of high and low settings of each factor, making the analysis outcome not as rigorous as it should be. Nevertheless, as long as the user is aware of those limitations and interprets the results of the analysis accordingly, we believe the data analysis methods are useful additions to the engineer's tool box, as illustrated by several other applications [24, 51, 52] scheduled for presentation at this conference.

9. CONCLUDING REMARKS

An NDE data analysis methodology using ANOVA, tolerance intervals, and design of experiments approaches for estimating reliability of weld flaw detection, location, and sizing, has been presented with examples based on the PVRC 1968 round-robin testing program of ultrasonic detection of implanted flaws in thick-section steel-plate weld specimens.

To take advantage of this new methodology, we have also demonstrated that it is feasible to implement the analysis capability in a web-based information system such that one can accelerate the feedback loop between the field inspectors of a structure or component for critical flaws by nondestructive evaluation (NDE) and the office engineers who do the damage assessment and recommendations for field action to prevent failure. For example, field inspection data of critical flaws can be transmitted to the office instantly via the internet, and the office engineer with a computer database of equipment geometry, material properties, past loading/deformation histories, and potential future loadings, can process the NDE data as input to a damage assessment model to simulate the equipment performance under a variety of loading conditions until its failure. Results of such simulations can be combined with engineering judgment to produce and transmit to the field specific recommendations, with uncertainty estimation, for operational decision making.

10. ACKNOWLEDGMENTS

11. REFERENCES

[1] Hedden, O. F., 1981, PVRC Intercommittee Correspondence to J. T. Fong, March 16, 1981. Available from PVRC, 345 E 47 St., New York, NY, 10017 (1981).

[2] Berger, H., and Smith, J. H., eds., 1976, Consideration of Fracture Mechanics Analysis and Defect Dimension Measurement Assessment for the Trans-Alaska Oil Pipeline Girth Welds, Nat. Bur. of Stand. Report to U.S. Dept. of Transportation, NBSIR 76-1154, Vols. I & II, Oct. 1976. Washington DC: Government Printing Office (1976).

[3] Lewis, W. H., Sproat, W. H., Dodd, B. D., and Hamilton, J. M., 1978, "Reliability of Nondestructive Inspections - Final Report," Lockheed-Georgia Report No. SA-ALC/MME 76-6-38-1 for Air Force Contract No. F41608-76-D-A005 Covering Period from July 1974 to Dec. 1978. Kelly AFB, TX: San Antonio Air Logistics Center (1978).

[4] Johnson, D. P., Toomay, T. L., and Davis, C. S., 1979, "Estimation of the Defect Detection Probability for Ultrasonic Tests on Thick Section Steel Weldments," EPRI NP-991 Project 700-1 Technical Report, Feb. 1979. Palo Alto, CA: Electric Power Research Institute (1979).

[5] Adamonis, D. C., and Hughes, E. T., 1979, "Ultrasonic Evaluation and Sectioning of PVRC Plate Weld Specimen 201," WRC Bulletin 252, September 1979.

[6] Ruescher, E. H., and Graber, H. C., 1980, "Analysis of the Radiographic Evaluation of PVRC Weld Specimens 155, 202, 203, and 251J," WRC Bulletin 259, June 1980. New York, NY: Welding Research Council (1980).

[7] Yukawa, S., 1981, "Sectioning Data of PVRC Plate-Weld Specimen 251J," Progress Report to PVRC Subcom. on NDE of Pressure Comp., Jan. 15, 1981. Available from PVRC, 345 E 47 St., New York, NY, 10017 (1981).

[8] Yukawa, S., 1983, "Sectioning of PVRC NDE Specimen 251J," Progress Report to PVRC Subcom. on NDE of Pressure Comp., May 25, 1983. Available from PVRC, 345 E. 47 St., New York, NY 10017 (1983).

[9] Yukawa, S., 1984, "Sectioning Data of PVRC Plate-Weld Specimen 251J," Progress Report to PVRC Subcom. on NDE of Pressure Comp., May 23, 1984. Avail. from PVRC, 345 E 47 St., New York, NY 10017 (1984).

[10] Yukawa, S., 1986, "Sectioning Results for PVRC Plate Weld Specimen 251J," Proc. June 1983 Portland Symp. on NDE Reliability Through Round Robin Testing, ASME Spec. Pub. NDE-1. New York, NY: Am. Soc. of Mech. Engineers (1986).

[11] Fong, J. T., 1982, "A 3-dimensional Computer-Aided Graphical Representation of the Sectioning Data of Flaw A in PVRC Plate-Weld Specimen 251J," Nat. Bur. of Stands. Report to PVRC Subcom. on NDE of Pressure Com, Sep. 28, 1982. Available from PVRC, 345 E 47 St, New York, NY 10017 (1982).

[12] Fong, J. T., 1986, "Analysis of PVRC 251J Sectioning Data (1984) and Round Robin Ultrasonic Test Data (1968) for Estimating Reliability of Flaw Fabrication and NDE Procedure," Nat. Bur. of Stands. Report (Final) to PVRC Subcom. on NDE of Pressure Comp., Feb. 10, 1986. Available from PVRC, 345 E 47 St., NY, NY, 10017, or, fong@nist.gov (1986).

[13] Fong, J. T., 1986, "Analysis of Sectioning Data of PVRC 251-J for Estimating Flaw Fabrication Reliability," Proc. ASME PVP Symp. on NDE Reliability through Round Rcbin Testing, June 1986, Portland, OR., J. T. Fong, L. Mordfin, O. F. Hedden, and S. H. Bush, eds., ASME Special Publication NDE-1, pp. 89-116. New York, NY: Am. Soc. of Mech. Engineers (1986).

[14] Fong, J. T., and Filliben, J. J., 1986, "A Data Analysis Methodology for Estimating NDE Reliability from Round Robin Ultrasonic Testing Data," Proc. ASME PVP Symp. on NDE Reliability through Round Rcbin Testing, June 1986, Portland, OR., J. T. Fong, L. Mordfin, O. F. Hedden, and S. H. Bush, eds., ASME Special Publication NDE-1, pp. 201-240. New York, NY: Am. Soc. of Mech. Engineers (1986).

[15] Comm. EC, 1979, "Destructive Examination of the PVRC plates Nos. 50-52, 51-53, & 204," EUR-6371, Vol. III, published by the Commission of European Communities. Luxembourg: Office for Official Publications of the European Communities (1979).

[16] Comm. EC, 1979, "Analysis Scheme of the PISC trials results," EUR - 6371, Vol. IV, published by the Commission of European Communities. Luxembourg: Office for Official Publications of the European Communities (1979).

[17] Comm. EU, 1985, "First Evaluation of the PISC-II Trials Results," PISC-II Report No. 5, June 1985. Luxembourg: Office for Official Publications of the European Communities (1985).

[18] Watkins, B., Ervine, R. W., and Cowburn, K. J., 1982, "The UKAEA Defect Detection Trials (DDT)," Presented at the UKAEA DDT Symposium, Birchwood, Warrington, 7-8 Oct. 1982. Available from UKAEA Risley Nuclear Laboratories, Risley, Warrington, U.K. (1982).

[19] Saiga, Y., 1983, "Sectioning of PVRC Test Plate 202," Japan PVRC Rep. to U.S. PVRC Subcom. on NDE of Press. Comp., Oct. 12, 1983. Available from PVRC, 345 E 47 St., NY, NY 10017 (1983).

[20] Saiga, Y., 1984, "Sectioning of PVRC Specimen 203," Japan PVRC Rep. to U.S. PVRC Subcom. on NDE of Press. Comp., May 23, 1984. Available from PVRC, 345 E 47 St., NY, NY 10017 (1984).

[21] Saiga, Y., 1984, "Analysis of the Ultrasonic Examinations of PVRC Weld Specimen 202," Japan PVRC Draft Rep. to U.S. PVRC Subcom. on NDE of Press. Comp., Jan. 23, 1984. Available from PVRC, 345 E 47 St., New York, NY, 10017 (1984).

[22] Saiga, Y., 1984, "Analysis of the Ultrasonic Examinations of PVRC Weld Specimen 203," Japan PVRC Draft Rep. to U.S. PVRC Subcom. on NDE of Press. Comp., May 23, 1984. Available from PVRC, 345 E 47 St., New York, NY, 10017 (1984).

[23] Bush, S. H., and Hedden, O. F., 2007, "Flaw Detection, Location, and Sizing," Proc. Digest of ASME PVP Symposium on Engineering Safety, Applied Mechanics, and Non-destructive Evaluatioon (NDE), in honor of Dr. Spencer H. Bush (1920-2005), July 2007, San Antonio, TX, pp. 153-156. Published by Stanford Mechanics Alumni Club, 104 King Farm Blvd, Suite C308, Rockville, MD 20850 (2007).

[24] Fong, J. T., Filliben, J. J., Heckert, N. A., Ranson, W. F., Vachon, R. I., and Marcal, P. V., 2008, "Structural Aging Monitoring via Web-based Nondestructive Evaluation Technology," Proc. ASME PVP Conference, July 27-31, 2008, Chicago, IL, Paper No. PVP2008-61607. New York, NY: ASME (2008).

[25] Filliben, J. J., 1984, DATAPLOT - Introduction and Overview, A special publication of the National

Bureau of Standards, SP667, 119pp. Washington, DC: Government Printing Office (1984).

[26] Filliben, J. J., and Heckert, N. A., 2002, DATAPLOT: A Statistical Data Analysis Software System, a public domain software released by NIST, Gaith, MD 20899, http://www.itl.nist.gov/div898/software/dataplot.html (2002).

[27] Gillette, C. L., and Smedley, W. C., 1968, "Ultrasonic Examination of Weld Seams in PVRC Test Blocks Welded at Chattanooga," Combustion Engineering Report E-7322 to PVRC, Oct. 14, 1968 (Ref. SCPVS-68-252). Available from PVRC, 345 East 47th Street, New York, NY 10017 (1968).

[28] White, J. K., 1968, "Inspection of Weld Specimens at Combustion Engineering, July 8, 9, & 10, 1968," Westinghouse Res. Lab. Progress Report to PVRC, Sep. 27, 1968 (Ref. SCPVS-668-105). Available from PVRC, 345 E. 47th St., New York, NY 10017 (1968).

[29] Buchanan, R. A., 1976, "Analysis of the NDE of PVRC Plate-Weld Specimen 251J," WRC Bulletin 221, Nov. 1976. New York, NY 10017: Welding Research Council (1976).

[30] Hedden, O. F., 1984, PVRC Intercommittee Correspondence to J. T. Fong, Feb. 17, 1984.

[31] Hedden, O. F., 1984, PVRC Intercommittee Correspondence to J. T. Fong, March 7, 1984.

[32] Hedden, O. F., 1985, PVRC Intercommittee Correspondence to J. T. Fong, Oct. 28, 1985.

[33] Nelson, P. R., Coffin, M., and Copeland, K. A. F., 2003, Introductory Statistics for Engineering Experimentation. Elsevier (2003)

[34] Filliben, J. J., 1969, "Simple and Robust Linear Estimation of the Location Parameter of a Symmetric Distribution," Ph.D. dissertation, Princeton University (1969).

[35] Filliben, J. J., 1975, "The Probability Plot Correlation Coefficient Test for Normality," Technometrics, Vol. 17, No. 1, pp. 111-117 (1975).

[36] Daniels, H. E., 1939, "The estimation of components of variance," J. Roy. Stat. Soc., Supp., Vol. 6, pp. 186-197 (1939).

[37] Eisenhart, C., 1947, "The Assumptions Underlying the Analysis of Variance," Biometrics, Vol. 3, No. 1, pp. 1-21 (1947).

[38] Crump, S. L., 1949, "The Estimation of Variance Components in Analysis of Variance," Biometrics Bulletin, Vol. 2, pp. 7-11 (1946).

[39] Hendricks, W. A., 1951, "Variance Components as a tool for the Analysis of Sample Data," Biometrics, Vol. 7, No. 1, pp. 97-101 (1951).

[40] Mandel, J., 1971, "Repeatability and Reproducibility," Materials Research and Standards, Vol. 11, No. 8, pp. 8-16 (1971).

[41] Mandel, J., 1977, "The Analysis of Interlaboratory Test Data," ASTM Standardization News, Vol. 5, No. 3, pp. 17-20 (1977).

[42] Draper, N. E., and Smith, H., 1981, Applied Regression Analysis, Second Edition. Wiley (1981).

[43] Proschan, F., 1953, "Confidence and Tolerance Intervals for the Normal Distribution," J. Am. Stat. Assoc., Vol. 48, pp. 550-564 (1953).

[44] Natrella, M. G., 1963, Experimental Statistics, NBS Handbook 91, 1963 edition (reprinted October 1966 with corrections). U. S. Government Printing Office, Washington, DC (1966).

[45] Beyer, W. H., ed., 1966, CRC Handbook of Tables for Probability and Statistics. Cleveland, OH: Chemical Rubber Co. (1966).

[46] Box, G. E., Hunter, W. G., and Hunter, J. S., 1978, Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. Wiley (1978).

[47] Montgomery, D. C., 2000, Design and Analysis of Experiments, 5th ed. Wiley (2000).

[48] Croarkin, C., Guthrie, W., Heckert, N. A., Filliben, J. J., Tobias, P., Prins, J., Zey, C., Hembree, B., and Trutna, eds., 2003, NIST/SEMATECH e-Handbook of Statistical Methods, Chap. 5 on Process Improvement, Produced jointly by the Statistical Eng. Div. of the Nat. Inst. of Stand. & Tech., Gaithersburg, MD, and the Statistical Methods Group of SEMITECH, Austin, TX, http://www.itl.nist.gov/div898/handbook/ (2003).

[49] Fong, J. T., Filliben, J. J., Heckert, N. A., and deWit, R., 2008, "Design of Experiments Approach to Verification and Uncertainty Estimation of Simulations based on Finite Element Method," to appear in Proc. June 22-25, 2008 Conf. Amer. Soc. for Eng. Education, Pittsburgh, PA, Paper No. AC2008-2725 (2008).

[50] Behravesh, M. M., and Dau, G. J., 1986, "Requalification of NDE Personnel for BWR IGSCC Detection," A Status Report to PVRC Subcom. on NDE of Pressure Components, Jan. 28, 1986, Available from PVRC, 345 E. 47th St., New York, NY 10017 (1986).

[51] Chao, Y., Lam, P. S., Fong, J. T., deWit, R., and Filliben, J. J., 2008, "A New Approach to Assessing the Reliability of Applying Laboratory Fracture Toughness Test Data to Full-Scale Structures," Proc. ASME PVP Conference, July 27-31, 2008, Chicago, IL, Paper No. PVP2008-61584. New York, NY: ASME (2008).

[52] Interrante, C. G., Fong, J. T., Filliben, J. J., and Heckert, N. A., 2008, "Uncertainty Estimate of Charpy Data using a 5-factor 9-run Design of Experiments," Proc. ASME PVP Conference, July 27-31, 2008, Chicago, IL, Paper No. PVP2008-61565. New York, NY: ASME (2008).

Table 1. A listing of **TFN**$_{P,n}$, the two-sided **T**olerance limit **F**actors for **N**ormal distributions at coverage P and sample size n, for two specific probabilities, $\gamma = 0.75$ and 0.90, such that the probability is γ for at least a proportion P of the distribution to be included between $X \pm$ **TFN**$_{P,n} * s$, where X and s are estimates of the mean and the standard deviation computed from a sample size n (after Proschan [43], Natrella [44], Beyer [45], Nelson et al [33], etc.).

	$\gamma = 0.75$					$\gamma = 0.90$				
P \\ n	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
2	4.498	6.301	7.414	9.531	11.920	11.407	15.978	18.800	24.167	30.227
3	2.501	3.538	4.187	5.431	6.844	4.132	5.847	6.919	8.974	11.309
4	2.035	2.892	3.431	4.471	5.657	2.932	4.166	4.943	6.440	8.149
5	1.825	2.599	3.088	4.033	5.117	2.454	3.494	4.152	5.423	6.879
6	1.704	2.429	2.889	3.779	4.802	2.196	3.131	3.723	4.870	6.188
7	1.624	2.318	2.757	3.611	4.593	2.034	2.902	3.452	4.521	5.750
8	1.568	2.238	2.663	3.491	4.444	1.921	2.743	3.264	4.278	5.446
9	1.525	2.178	2.593	3.400	4.330	1.839	2.626	3.125	4.098	5.220
10	1.492	2.131	2.537	3.328	4.241	1.775	2.535	3.018	3.959	5.046
11	1.465	2.093	2.493	3.271	4.169	1.724	2.463	2.933	3.849	4.906
12	1.443	2.062	2.456	3.223	4.110	1.683	2.404	2.863	3.758	4.792
13	1.425	2.036	2.424	3.183	4.059	1.648	2.355	2.805	3.682	4.697
14	1.409	2.013	2.398	3.148	4.016	1.619	2.314	2.756	3.618	4.615
15	1.395	1.994	2.375	3.118	3.979	1.594	2.278	2.713	3.562	4.545
16	1.383	1.977	2.355	3.092	3.946	1.572	2.246	2.676	3.514	4.484
17	1.372	1.962	2.337	3.069	3.917	1.552	2.219	2.643	3.471	4.430
18	1.363	1.948	2.321	3.048	3.891	1.535	2.194	2.614	3.433	4.382
19	1.355	1.936	2.307	3.030	3.867	1.520	2.172	2.588	3.399	4.339
20	1.347	1.925	2.294	3.013	3.846	1.506	2.152	2.564	3.368	4.300
21	1.340	1.915	2.282	2.998	3.827	1.493	2.135	2.543	3.340	4.264
22	1.334	1.906	2.271	2.984	3.809	1.482	2.118	2.524	3.315	4.232
23	1.328	1.898	2.261	2.971	3.793	1.471	2.103	2.506	3.292	4.203
24	1.322	1.891	2.252	2.959	3.778	1.462	2.089	2.489	3.270	4.176
25	1.317	1.883	2.244	2.948	3.764	1.453	2.077	2.474	3.251	4.151
26	1.313	1.877	2.236	2.938	3.751	1.444	2.065	2.460	3.232	4.127
27	1.309	1.871	2.229	2.929	3.740	1.437	2.054	2.447	3.215	4.106

Table 2. A listing of **TFN**$_{P,n}$, the two-sided **T**olerance limit **F**actors for **N**ormal distributions at coverage P and sample size n, for two specific probabilities, $\gamma = 0.95$ and 0.99, such that the probability is γ for at least a proportion P of the distribution to be included between $X \pm$ **TFN**$_{P,n} * s$, where X and s are estimates of the mean and the standard deviation computed from a sample size n (after Proschan [43], Natrella [44], Beyer [45], Nelson et al [33], etc.).

n \ P	$\gamma = 0.95$					$\gamma = 0.99$				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
2	22.858	32.019	37.674	48.430	60.573	114.363	160.193	188.491	242.300	303.054
3	5.922	8.380	9.916	12.861	16.208	13.378	18.930	22.401	29.055	36.616
4	3.779	5.369	6.370	8.299	10.502	6.614	9.398	11.150	14.527	18.383
5	3.002	4.275	5.079	6.634	8.415	4.643	6.612	7.855	10.260	13.015
6	2.604	3.712	4.414	5.775	7.337	3.743	5.337	6.345	8.301	10.548
7	2.361	3.369	4.007	5.248	6.676	3.233	4.613	5.488	7.187	9.142
8	2.197	3.136	3.732	4.891	6.226	2.905	4.147	4.936	6.468	8.234
9	2.078	2.967	3.532	4.631	5.899	2.677	3.822	4.550	5.966	7.600
10	1.987	2.839	3.379	4.433	5.649	2.508	3.582	4.265	5.594	7.129
11	1.916	2.737	3.259	4.277	5.452	2.378	3.397	4.045	5.308	6.766
12	1.858	2.655	3.162	4.150	5.291	2.274	3.250	3.870	5.079	6.477
13	1.810	2.587	3.081	4.044	5.158	2.190	3.130	3.727	4.893	6.240
14	1.770	2.529	3.012	3.955	5.045	2.120	3.029	3.608	4.737	6.043
15	1.735	2.480	2.954	3.878	4.949	2.060	2.945	3.507	4.605	5.876
16	1.705	2.437	2.903	3.812	4.865	2.009	2.872	3.421	4.492	5.732
17	1.679	2.400	2.858	3.754	4.791	1.965	2.808	3.345	4.393	5.607
18	1.655	2.366	2.819	3.702	4.725	1.926	2.753	3.279	4.307	5.497
19	1.635	2.337	2.784	3.656	4.667	1.891	2.703	3.221	4.230	5.399
20	1.616	2.310	2.752	3.615	4.614	1.860	2.659	3.168	4.161	5.312
21	1.599	2.286	2.723	3.577	4.567	1.833	2.620	3.121	4.100	5.234
22	1.584	2.264	2.697	3.543	4.523	1.808	2.584	3.078	4.044	5.163
23	1.570	2.244	2.673	3.512	4.484	1.785	2.551	3.040	3.993	5.098
24	1.557	2.225	2.651	3.483	4.447	1.764	2.522	3.004	3.947	5.039
25	1.545	2.208	2.631	3.457	4.413	1.745	2.494	2.972	3.904	4.985
26	1.534	2.193	2.612	3.432	4.382	1.727	2.469	2.941	3.865	4.935
27	1.523	2.178	2.595	3.409	4.353	1.711	2.446	2.914	3.828	4.888

Table 3.   Team A (PVRC 1968 Procedure) UT Data as corrected by Hedden [31] for analysis by Fong et al [11-12].

Team A (1968 Procedure) UT Data (Rev. by Hedden, 1984-03-07)

File Name: [PVRC] UT68AY.DAT

INDICA.	DACMAX.	X1UAY	X2UAY	Y1UAY	Y2UAY	Z1UAY	Z2UAY
1	50	24.79	24.81	30.60	32.60	7.99	8.01
2	80	25.59	25.61	28.95	31.55	9.35	10.45
3	50	24.59	24.61	29.05	32.05	10.30	10.65
4	50	26.49	26.51	29.00	32.00	4.49	4.51
5	80	25.89	25.91	14.80	19.80	4.65	5.75
6	100	25.79	25.81	16.80	17.80	8.29	8.31
7	50	25.79	25.81	15.30	19.30	7.84	7.86
8	110	26.59	26.61	1.50	4.00	8.14	8.16
9	50	24.79	24.81	1.80	3.00	6.24	6.26
10	25	25.59	25.61	4.30	5.80	3.00	3.60
11	40	25.59	25.61	2.30	4.80	3.40	3.95
12	110	25.59	25.61	1.80	3.30	3.00	4.05
13	80	26.59	26.61	3.00	4.50	8.24	8.26
14	120	24.79	24.81	2.30	4.30	6.34	6.36
15	100	26.79	26.81	15.00	18.00	1.99	2.01
16	60	25.89	25.91	16.20	20.50	4.75	5.85
17	50	25.79	25.81	17.50	19.00	7.74	7.76
18	200	24.60	25.60	28.50	31.50	9.45	10.55

Table 4. Team B (PVRC 1968 Procedure) UT Data (uncorrected) for analysis by Fong et al [11-12].

Team B (1968 Procedure) UT Data (Uncorrected)

File Name: [PVRC] UT68B.DAT

INDICA.	DACMAX.	X1UTC	X2UTC	Y1UTC	Y2UTC	Z1UTC	Z2UTC
1	25	25.20	25.70	2.00	6.00	5.50	6.60
2	30	25.20	25.70	16.00	18.50	9.00	9.50
3	25	26.30	26.70	16.30	22.30	6.50	7.00
4	30	26.30	26.60	28.30	32.20	9.00	9.50

Table 5. Team C (PVRC 1968 Procedure) UT Data (uncorrected) for analysis by Fong et al [11-12].

Team C (1968 Procedure) UT Data (Uncorrected)

File Name: [PVRC] UT68C.DAT

INDICA.	DACMAX.	X1UTC	X2UTC	Y1UTC	Y2UTC	Z1UTC	Z2UTC
1	100	24.98	25.02	15.50	18.50	0.98	1.02
2	40	24.98	25.02	30.50	33.75	5.48	5.52
3	75	24.98	25.02	1.00	4.50	9.98	10.02
4	40	25.98	26.02	1.75	2.50	2.73	2.77
5	60	25.98	26.02	16.00	21.00	8.23	8.27
6	25	24.98	25.02	1.50	4.50	0.98	1.02
7	40	26.48	26.52	30.00	34.25	2.73	2.77
8	40	25.48	25.52	14.50	20.00	5.48	5.52
9	45	25.98	26.02	1.00	6.00	8.23	8.27
10	75	24.98	25.02	30.00	34.00	9.98	10.02
11	11	25.98	26.02	31.98	32.02	0.98	1.02
12	9	25.98	26.02	16.98	17.02	2.73	2.77
13	8	25.98	26.02	1.98	2.02	5.48	5.52
14	9	25.98	26.02	16.98	17.02	9.98	10.02