Proceedings of the ASME 2018 Pressure Vessels and Piping Division Conference
PVP2018
July 15-20, 2018, Prague, Czech Republic

Preprint

PVP2018-84771

THREE APPROACHES TO QUANTIFICATION OF NDE UNCERTAINTY AND A DETAILED EXPOSITION OF THE EXPERT PANEL APPROACH USING THE SHEFFIELD ELICITATION FRAMEWORK

Jeffrey T. Fong
National Institute of Standards & Technology
Gaithersburg MD 20899 U.S.A.
fong@nist.gov

N. Alan Heckert
National Inst. of Stand. & Tech.
Gaithersburg MD 20899 U.S.A.
alan.heckert@nist.gov

James J. Filliben
National Institute of Standards & Technology
Gaithersburg MD 20899 U.S.A.
filliben@nist.gov

Steven R. Doctor
Independent Consultant (*)
Richland, WA 99352 U.S.A.
doctor4244@aol.com

ABSTRACT

The ASME Boiler & Pressure Vessel Code Section XI Committee is currently developing a new Division 2 nuclear code entitled the "Reliability and Integrity Management (RIM) program," with which one is able to arrive at a risk-informed, non-destructive examination (NDE)-based engineering maintenance decision by estimating and managing all uncertainties for the entire life cycle including design, material selection, degradation processes, operation and NDE. This paper focuses on the uncertainty of the NDE methods employed for preservice and inservice inspections due to a large number of factors such as the NDE equipment type and age, the operator's level and years of experience, the angle of probe, the flaw type, etc. In this paper, we describe three approaches with which uncertainty in NDE-risk-informed decision making can be quantified: (1) A regression model approach in analyzing round-robin experimental data such as the 1981-82 Piping Inspection Round Robin (PIRR), the 1986 Mini-Round Robin (MRR) on intergranular stress corrosion cracking (IGSCC) detection and sizing, and the 1989-90 international Programme for the Inspection of Steel Components III-Austenitic Steel Testing (PISC-AST). (2) A statistical design of experiments approach. (3) An expert knowledge elicitation approach. Based on a 2003 Pacific Northwest National Laboratory (PNNL) report by Heasler and Doctor (NUREG/CR-6795), we observe that the first approach utilized round robin studies that gave NDE uncertainty information on the state of the art of the NDE technology employed from the early 1980s to the early 1990s. This approach is very time-consuming and expensive to implement. The second approach is based on a design-of-experiments (DEX) of eight field inspection exercises for finding the length of a subsurface crack in a pressure vessel head using ultrasonic testing (UT), where five factors (operator's service experience, UT machine age, cable length, probe angle, and plastic shim thickness), were chosen to quantify the sizing uncertainty of the UT method. The DEX approach is also time-consuming and costly, but has the advantage that it can be tailored to a specific defect-detection and defect-sizing problem. The third approach using an expert panel is the most efficient and least costly approach. Using the crack length results of the second approach, we introduce in this paper how the expert panel approach can be implemented with the application of a software package named the Sheffield Elicitation Framework (SHELF). The crack length estimation with uncertainty results of the three approaches are compared and discussed. Significance and limitations of the three uncertainty quantification approaches to risk assessment of NDE-based engineering decisions are presented and discussed.

Keywords: Depth sizing; design of experiments; expert knowledge elicitation; flaw sizing; in-service inspection; intergranular stress corrosion cracking; logistic regression

(*) formerly with the Pacific Northwest National Laboratory.

model; nondestructive examination; pressure vessel; probability of detection; regression model; reliability; round robin studies; SHELF; ultrasonic testing; uncertainty quantification.

Disclaimer: Certain commercial equipment, materials, or software are identified in this paper in order to specify the computational procedure adequately. Such identification is not intended to imply endorsement by NIST, nor to imply that the equipment, materials, or software identified are necessarily the best available for the purpose.

1. INTRODUCTION

Uncertainty in quantitative results of nondestructive examination (NDE) techniques such as radiographic testing (RT) and ultrasonic testing (UT), as applied to the detection, sizing, and location of flaws in a structure or component has been of interest to engineers, scientists, and the public for a long time, simply because the results of such techniques can help us answer two fundamental questions, namely

(1) Flaw-caused damage diagnosis: Within 95 % confidence bounds, where is the flaw, what kind and how big?

(2) Damage prognosis: What shall we do and how much will mitigation cost to prolong the life of the aging structure?

In the mid-1970s, as reported by Adamonis and Hughes [1] and Hedden [2], the Welding Research Council, through its research arm, the Pressure Vessel Research Committee (PVRC), embarked on a major investigation of NDE reliability by developing a series of round robin UT examinations on 12 thick-section welded steel plate specimens containing carefully designed flaws (see Fig. 1) that were implanted and inspected by qualified teams using ASME and other industry standards and procedures then-approved by PVRC.

Unfortunately, the results of the PVRC NDE reliability research program were inconclusive for three principal reasons:

(1) The process of making weld specimens containing implanted flaws was imperfect because it used graphite sheets as flaws between hot welding passes that distorted the flaw lengths with amplification factors (AF) as large as 2.7 (see Refs. [3-5]) and estimated 95 % upper confidence limit of AF as large as 4.2 (as shown in Table 1 and Refs. [4, 5]).

(2) The lack of a statistically-sound design of experiments.

(3) The lack of an adequate number of qualified teams to participate for arriving at a statistically-sound basis for engineering judgment (only three UT teams were available).

Table 1. List of Implanted Flaws in PVRC Specimen 251J and their 95 % Upper Confid. Limits of Amplification Factors (AF) where AF = True Length / Initial Length of Implanted Flaw.

Case	Type of Flaws	Upper Limit of *AF*
5	All flaw types	**3.5**
2	Cross Cracks	**3.8**
3	Longitudinal Cracks	**4.2**
4	Slag Inclusions	**2.0**

Lessons learned from the PVRC NDE Reliability Program of the 1970s, and the PISC I Program (1975-80) that used some of the same PVRC samples and then became an international program by having European teams conduct inspections, led to new investigations in the early 1980s using the "Performance Demonstration Approach," modeled after the process specified by ASME for welding procedure development.

This performance demonstration process was already being pursued at the Pacific Northwest National Laboratory (PNNL), Richland, WA, in 1981 based on PNNL laboratory testing (see Doctor [6]) and the PVRC and PISC I results. In addition, the Nuclear Regulatory Commission (NRC) initiated performance demonstration through Inspection and Enforcement Bulletins IEB 82-03 and 83-02.

At the International Symposium on Reliability of Reactor Pressure Components, International Atomic Energy Agency (IAEA), Stuttgart, Germany, March 21-25, 1983 (see a historical paper by Bush and Hedden [7]), the late Dr. Spencer Bush had sufficient confidence in the success of NDE reliability studies to declare the following visionary statement:

> *". . . 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."*

This rather optimistic scenario was based primarily on the assumption that the problem of quantifying the uncertainty in NDE reliability had been solved. In 2008, a web-based field-office-field NDE data transmission and analysis methodology was developed by Fong, Hedden, Filliben, and Heckert [8] to create a fast link between field results of Monitoring And NDE (MANDE) and fracture mechanics-based remaining life predictive models for on-site maintenance decision making. A recent review including Refs. [9-12] addressing the state of the science (uncertainty quantification) and art (engineering judgment) of NDE reliability led us to conclude that the 1983 Bush statement was premature, and rather a work in progress.

To show that the NDE uncertainty quantification (UQ) problem is far from being completely solved, we review in this paper the state of science and the art of NDE reliability by examining three NDE UQ approaches. In Section 2, we introduce **Approach-1** (Regression Models) with the presentation of NDE data analysis results from four major studies.

In Section 3, we present **Approach-2** (Design of Experiments) with the analysis of the results of a field UT

experiment on the examination of a single crack by 8 teams with five factors.

In Section 4, we introduce **Approach-3** (Knowledge Elicitation) with the Sheffield Elicitation Framework (SHELF). An example for conducting an expert panel involving four UT experts to give judgments on the length of a subsurface crack in a pressure vessel head appears in Section 5.

A discussion of the pros and cons of the three NDE uncertainty quantification approaches is given in Section 6. Some concluding remarks and a list of references appear in Sections 7 and 8, respectively.

2. APPROACH-1: REGRESSION MODELS

To quantify NDE uncertainty, we need well-characterized NDE data, validated true state flaw data, and state-of-the-art statistical analysis methods. During the last sixty years, a great deal of NDE data have been generated, some available in the open literature but most remaining unpublished because of the confidential nature of the inspection work. Fortunately, in the applied statistics literature, we could identify three uncertainty quantification (UQ) approaches for analyzing data of diverse origins, namely (**Approach-1**) regression models, (**Approach-2**) design of experiments, and (**Approach-3**) subjective probability-based knowledge elicitation.

In this section, we will review four major NDE-UT (ultrasonic testing) databases using **Approach-1**, and summarize the state of UQ of three NDE parameters of interest, namely (a) probability of detection (POD), (b) flaw length sizing, and (c) flaw depth sizing. The four NDE databases are:

(1) The 1981-82 Piping Inspection Round Robin (PIRR) [13] by the Pacific Northwest National Laboratory (PNNL) wrought stainless steel (WSS) with IGSCC and thermal fatigue cracks (TFC).

(2) The 1985 Mini-Round Robin (MRR) [14] by PNNL WSS with only IGSCC.

(3) The 1989-1990 Programme for Inspection of Steel Components - Phase III - Austenitic Steel Test (PISC-AST) [15] by the Joint Research Centre of the Commission of the European Communities WSS with EDM, IGSCC, mechanical fatigue cracks (MFC), and TFC.

(4) The 1995[#]-2015 unpublished EPRI-Performance Demonstration Initiative (PDI) [16] on dissimilar metal weldments (DMW) with simulated primary water stress corrosion cracks (PWSCC). *[#]Note-1: The year 1995 is an estimate of the first year of the EPRI-PDI program that began in the mid-1990's and ran until 2015 according to Ref. [16].*

Details of the PIRR, MRR, and PISC-AST programs that generated their data are given in Tables 2, 3, and 4 (after [17]) where all of the inspections were conducted from the outside surface of the piping. A less informative summary of the EPRI-PDI program is given in Table 5 (after [16]) where some of the data is for outside surface inspections and some for inside surface inspections.

A key difference between the results of the first three databases (PIRR, MRR, and PISC-AST) and that of the fourth one (EPRI-PDI), is that the former lumped all cracks as a single type[##], and the latter divided the cracks into several distinct categories before analysis because DMW's have been found to experience PWSCC degradation in plants with both axial and circumferential orientations. From the UQ point of view, the results of the latter (EPRI-PDI) are better and more informative. *[##]Note-2: The interest was in structurally significant flaws in WSS that had been detected in service so the type of cracks used in these studies were dominantly circumferential flaws although some flaws have axial elements. Circumferential flaws may lead to guillotine failure but axial flaws will probably lead to leaking.*

Let us present our findings in two stages, with the first stage being a review of 9 selected NDE uncertainty plots from the four databases, and the second stage, a comparison of the four sets of NDE UQ results.

In Ref. [17], we found in one of many PIRR POD plots of the 553 inspections (45 flaws by 7 teams) with 95% confidence bounds a significant result, namely, for a flaw depth of 14.3 mm, the lower 95 % confidence limit for POD is 85 %. In a PISC-AST POD plot of the 133 inspections (26 flaws/23 teams) with 95 % confidence bounds [17], we found a comparable result, namely, for a flaw depth of 9.7 mm, the lower 95 % confidence

Table 2. Summary of First 3 NDE databases reported in [17].

	PIRR	MRR	PISC-AST
No. of Inspections	553	309	133
No. of Teams	7	15	23
No. of Assemblies	86	20	6
Ave. Wall Thickness, mm	14	14	21
Flaw depth, mm			
Min	0.33	0.83	0.40
Median	2.41	4.78	4.50
Max	6.83	11.44	14.10
Flaw Length, mm			
Min	3.05	3.30	0.52
Median	26.42	21.59	46.39
Max	59.19	130.80	108.20
Total No. of Flaws	**45**	**15**	**26**

Table 3. Flaw Lengths of First 3 NDE databases (after [17]).

	PIRR	MRR	PISC-AST
No. of Teams	6	14	23
No. of Flaws	36	13	26
No. of Tests	267	123	371
Ave Flaw Length, mm	**27.58**	**27.02**	**52.49**

Table 4. Flaw Depths of First 3 NDE databases (after [17]).

	PIRR	MRR	PISC-AST
No. of Teams	6	8	23
No. of Flaws	36	10	26
No. of Tests	267	80	374
Ave Flaw Depth, mm	**2.78**	**4.87**	**4.93**

limit for POD is also 85 %. That showed an improvement in UT detection capability, because a detectable depth of 9.7 mm is significantly better than that of the PIRR POD-based capability, i.e., 14.3 mm. That is interesting but not surprising, since the techniques, procedures and training between the early baseline of the PIRR and that of the PISC-AST were night and day different with the latter having nearly 10 years of learning from the round robin testing and field experience.

Again in Ref. [17], we found in one of many POD plots of the combined PIRR, MRR and PISC-AST data (995 inspections, 86 flaws, 45 teams) with 95 % confidence bounds a significant result, namely, for a flaw depth of 12.0 mm, the lower 95 % confidence limit for POD is also 85 %. It is interesting to observe that for 85 % POD at 95 % confidence, the detectable flaw depth is 12 mm, which is about half-way between the results of a stand-alone PISC-AST (9.7 mm) and PIRR (14.3 mm) database.

In other words, lumping together two or more sets of NDE data reduces an opportunity to differentiate the uncertainty characteristics of the individual sets. By extension, when all three sets, the PIRR, the MRR, and the PISC-AST, lumped all flaws together in the data analysis, one failed to differentiate the uncertainty characteristics of the subsets of individual flaw types. That problem did not occur with the EPRI-PDI database, because they carefully separated out the flaw types, as shown in the latter part of this section.

Again in Ref. [17], we found in one of many length sizing plots of the three databases, i.e., PIRR, MRR, and PISC-AST, a specific plot, in which upper and lower confidence bounds were identified for each of the three databases for one to do a visual comparison. Clearly, the bounds for PIRR and MRR are too broad for one to make an assessment. So we choose to look at the PISC-AST result and pick one specific length measurement to illustrate its UQ characteristics. For example, if the measured length of a flaw is 100 mm (mean), then the true flaw length is 129 mm (mean), and the 95 % confidence bounds give us the measured bounds to be the interval, (67, 133), and the true bounds to be the interval, (93, 200). The percent error of the mean measured length is - 23 %, and the percent error of the 95 % upper limit of the measured length is - 34 %. No such UQ results exist for PIRR and MRR databases.

Table 5. Flaw Categories of EPRI-PDI database (after [16]).

		Model Development Dataset		
Category	Flaw Orientation	Attempts	Detections	Detection Rate
A(PF)	Axial	863	661	77%
	Circumferential	2358	2131	90%
B1(PF)	Axial	308	289	94%
	Circumferential	590	576	98%
C(PF,UD)	Axial	932	833	89%
	Circumferential	2375	2147	90%
C(PF,CD)	Axial	932	820	88%
	Circumferential	2375	2059	87%

Also in Ref [17], we searched for a specific depth sizing plot of the same three databases we just investigated earlier for length sizing. Unfortunately, no UQ results exist for the depth sizing of any of the three databases, namly PIRR, MRR, and PISC-AST.

We now change our focus from the three databases reported by Heasler and Doctor [17] to the EPRI-PDI database [16], which, as shown in Table 5, consists of eight mini-databases to be identified in Table 6 as follows:

Table 6. List of 8 EPRI-PDI mini-databases (after [16]).

Mini-database No.	Category	Flaw Orientation
1	A (PF)	Axial
2	A (PF)	Circumferential
3	B1 (PF)	Axial
4	B1 (PF)	Circumferential
5	C (PF, UD)	Axial
6	C (PF, UD)	Circumferential
7	C (PF, CD)	Axial
8	C (PF, CD)	Circumferential

In Ref. [16], we located the length sizing plots of the EPRI-PDI mini-database no. 4 (Category B1 (inside surface inspection of RPV nozzles) circumferential flaws) with 95 % confidence bounds. The UQ results are extremely good. For example, if the measured length is 0.50 normalized units (mean), the true length is 0.54 units (mean), and the percent error is only - 8 %. The percent error of the 95 % upper limit of the measured length is - 18 %, which is much less than that of the much older PISC-AST database, namely - 34 %. That is good news for the length sizing reliability component of the NDE engineering science, but not so good for the depth sizing reliability component because both the EPRI-PDI and the first three databases (PIRR, MRR, PISC-AST) failed to show any UQ results. We will say more on this in Section 6 (Discussion).

Let us turn our attention to the POD UQ results of the EPRI-PDI mini-databases. Again in Ref. [16], we located a specific POD plot of the EPRI-PDI mini-database no. 2 (Category A (Pressurized surge) circumferential flaws) with 95 % confidence bounds for two distinct sets of NDE data, namely, ALL DATA (in blue), and Reduced DATA (in red, less outliers). For example, working with ALL DATA, for a flaw size of 6.0 mm, the lower 95 % confidence limit for POD is 85 %.

Also in Ref. [16], we located another POD plot of the EPRI-PDI mini-database no. 3 (Category B1 (RPV nozzles) axial flaws) with 95 % confidence bounds again for two distinct data sets. If we work with the ALL Data set (in blue), we found that for a flaw size of 6.0 mm, the lower 95 % confidence limit for POD is again 85 %. This shows that the two databases, Nos. 2 and 3, gave similar UT uncertainty characteristics.

Unfortunately, such similarity does not exist for mini-database no. 8. (Category C (weld overlays), characterized data circumferential flaws) with 95 % confidence bounds again for

two distinct data sets. Working with the ALL DATA set, for a flaw size of 7.2 mm, the lower 95 % confidence limit for POD is found to be 85 % with a negative lower limit slope that implies a less probable POD will have a larger lower limit, a physical impossibility. This casts doubt on the POD UQ results of the mini-databases nos. 5 through 8 (Category C (weld overlays)), indicating that a new analysis methodology may be needed to remove that physically impossible negative slope.

3. APPROACH-2: DESIGN OF EXPERIMENTS (DEX)

The NDE uncertainty results of **Approach-1** (Regression Models) are global in nature in the sense that the 95 % confidence bounds were estimated from a very diverse set of data with no way to find out the individual uncertainty component of any one of many factors that may influence the outcome of a specific NDE measurement process aimed at detecting, sizing, and locating a unique type of flaw that the engineers believe to be most likely to exist in a specific pressurized component. To remedy this shortcoming, we need an alternative approach to quantify NDE uncertainty "in the small," or, local in nature. Such is our **Approach-2** based on the statistical theory of Design of EXperiments (DEX).

There is a vast literature in statistics on DEX (see, for instance, Refs. [18, 19, 20]), but very few on applications to NDE data (see, for instance, Refs. [8, 21, 22]), partly because most NDE data are confidential and seldom in the public domain. For a tutorial on an application of DEX to NDE data, see Fong, Hedden, Filliben and Heckert [8]. In this section, we will introduce **Approach-2** through the description of an NDE DEX example that appeared in Fong, et al. [8].

Our example begins with the knowledge that a subsurface crack has been detected and located in a weldment of a pressure vessel by a single UT inspection team, and it is requested to design and execute a UT experiment in order to establish a credible measurement of the crack length with a 95 % predictive upper limit for a safety assessment based on fracture mechanics and fatigue life modeling.

In Fig. 1, we show the details of an experimental pogram to fulfill such a request, where five factors, X1, X2, X3, X4, and X5, have been identified (k = 5), each with its low, center, and high values assigned. The response variable, Y1, has also been identified as the crack length. If the design is a so-called two-level full factorial experiment, we will need 2^5 , or, 32 UT teams to inspect the same weld, i.e., n = 32. It is, however, possible to reduce n , by making the design a two-level *fractional* factorial experiment, with n always a power of 2, such as 32, 16, 8, etc., but never less than k (=5). So we choose to conduct an experiment with 8 UT teams with their length measurements given by Y1 in a table in Fig. 2 as follows:

43.2, 54.6, 44.5, 67.3, 57.2, 53.3, 33.0, 63.5 mm, or,
1.70, 2.15, 1.75, 2.65, 2.25, 2.10, 1.30, 2.50 inches.

In Figs. 3 and 4, we show the key results of the DEX analysis where we learn from Fig. 3 that among the five factors, two are dominant, namely X1 (service year) and X4 (probe angle). In Fig. 4, we see that there are three non-trivial pair-interactions , namely X2-X4, X3-X5, and X1-X2. The size of the uncertainty, known as the "effect", of each factor with or without pair interactions, is given in a table listed in Fig. 5. From that table, we conclude that factors X1 and X4 could be singled out as dominant variables in a linear regression model with a 2-parameter response surface (Fig. 6) and a crack length distribution plot with 95 % predictive bounds given in Fig. 7.

As shown in Fig. 7, the NDE UQ result of **Approach-2 (DEX)** is given by the following crack length measurement:

Crack Length with Predictive Bounds = 52.1 (14.8) mm, or, 2.05 (0.58) inch. The 95 % upper limit = 66.9 mm (2.63 in.).

4. APPROACH-3: KNOWLEDGE ELICITATION/SHELF

Both **Approach-1** (Regression Models) for NDE UQ in the large and **Approach-2** (Design of Experiments) for UQ in the small are time consuming and expensive. **Approach 1** could take years and cost millions of dollars to complete, while **Approach-2**, months and hundreds of thousand dollars.

When something happens at a plant that requires a quick assessment with less than a month to conduct an inquiry and a budget of under one hundred thousand dollars, neither approach is feasible. For instance, as shown in our example in Section 3, **Approach-2** would require, for a problem with five factors, a minimum of 8 teams to complete a design of experiment exercise that would last at least six weeks including planning and data analysis.

Fortunately, there is an alternative to **Approach 2** that requires a minimum of two and a maximum of four (preferably three) teams to obtain a local NDE UQ result, and that is **Approach-3** (Expert Knowledge Elicitation).

So what is "expert knowledge elicitation (EKE)?" Before we answer this question, it is useful to give a brief history of how this new approach of UQ came into being. In 1967, Winkler [23] introduced the concept of quantifying judgment in a formal statistical framework, and in 1977, he and Murphy [24] applied the concept of "subjective probability" to weather forecasting. Two schools of aggregating individual expert opinions soon emerged, namely a formal model pioneered by Cooke [25] requiring 5-20 experts, and a behavioural model by O'Hagan [26-28] and Oakley and O'Hagan [29] requiring a minimum of only two experts. For our NDE UQ applications, we choose to adopt the behavioural aggregation model and apply the **Sh**effield **El**icitation **F**ramework (SHELF), its computer-assisted protocol (see Gosling [30]), as **Approach-3**.

As defined by O'Hagan [31], EKE is

". . . the process of representing the knowledge of one or more persons (experts) concerning an uncertain quantity as a probability distribution for that quantity."

As stated by Gosling [30, Section 4.1],

"*SHELF is an EKE protocol that provides a transparent and rigorous approach to capturing judgments from multiple experts. The synthesis of the experts' judgments is achieved through facilitated group discussion aiming to arrive at a consensus distribution using behavioural aggregation.*"

In Fig. 8, we show a flow chart of a typical SHELF exercise with 8 activities including the identification of experts, a facilitator, a recorder, and a quantity of interest, preparation of evidence dossier, training of experts, conducting of workshops, and the arrival at a consensus distribution. In Fig. 9, we show a SHELF-facilitated parametric fit of a typical response of an expert when asked to estimate an unknown quantity in terms of five "subjective probabilities" as shown in the following numerical example:

Subjective Probability	Value of the Unknown Quantity
0 %	(L = 0 , Lower Bound)
33 %	(T1 = 30 , Lower Tertile)
50 %	(M = 35 , Median)
67 %	(T2 = 55 , Upper Tertile)
100 %	(U = 100 , Upper Bound)

Note that the SHELF software recommends the use of either tertiles or quartiles (two points) with three more points (the median, the lower, and the upper bound) to form a 5-point representation of an expert's subjective probabilities, and gives the user a choice of one from several distributions to do a parametric fit as shown in the example given in Fig. 9. It turned out that the SHELF software happened to choose the Beta distribution for the parameter fit in Fig. 9 for the five opinion values furnished by the expert using the tertile-system. But the expert could have asked for an alternative one and the SHELF will respond with a new choice until the expert is satisfied.

To illustrate how a SHELF-based expert panel works, we develop in the next section an example based on an NDE crack length sizing UQ example given in **Approach 2.**

5. EXAMPLE OF AN EXPERT PANEL USING SHELF

Before identifying a quantity of interest (QoI), and recruiting experts, a facilitator, and a recorder, it is necessary to get organized and assign responsibilities for preparing documents prior to and during the convening of the expert panel. A minimum of four people need to be identified with responsibilities to prepare five documents as follows:

1. Client. The client is the person or a representative of an organization that requires the elicitation. His responsibility is to (a) prepare Elicitation Document ED-1 (Quantity of Interest, or, QoI), (b) prepare Elicitation Document ED-2 (Expert Enquiry Form), and (c) recruit a Coordinator.

2. Coordinator. The coordinator is an administrator who is knowledgeable about the QoI and is responsible for (a) preparing Elicitation Document ED-3 (Evidence Dossier for the QoI), (b) distributing ED-1, ED-2, and ED-3 to potential expert candidates, (c) selecting experts based on their ED-2's, (d) recruiting a Facilitator and a Recorder for a SHELF exercise, and (e) set up a date, time, and place for a workshop.

3. Facilitator. The facilitator manages the elicitation workshop, and is responsible for preparing Elicitation Document ED-4 (Context for Pre-Elicitation) during the convening of the first part of the workshop. ED-4 is required to contain the following items of information:

ED-4.01 Elicitation Title.
ED-4.02 Date and Start Time of Workshop Part 1.
ED-4.03 Attendance and Roles.
ED-4.04 Purpose of Elicitation.
ED-4.05 Orientation and Training.
ED-4.06 Participants' Expertise.
ED-4.07 Participants' Declaration of Interests.
ED-4.08 Strengths and Weaknesses of the Panel.
ED-4.09 Evidence
ED-4.10 Definitions
ED-4.11 End Time of Workshop Part 1 (Pre-Elicitation).
ED-4.12 Attachment-1: SHELF Expert Briefing.
ED-4.13 Attachment-2: Training presentation.
ED-4.14: Attachment-3: Workshop Part 2 Record Form.

4. Recorder. The recorder is proficient in using the SHELF computer software, which is written in a computer language named "R." The recorder takes the responsibility of completing the SHELF templates while the facilitator is managing the interaction between experts. More often, the recorder takes detailed notes during the workshop in order to complete the templates soon afterwards. The recorder runs the SHELF software and displays the distribution results during the second part of the workshop. The recorder is responsible for preparing Elicitation Document ED-5 with the following entries:

ED-5.01 Elicitation Title
ED-5.02 Date and Start Time of Workshop Part 2.
ED-5.03 Quantity of Interest.
ED-5.04 Anonymous Identity of Experts, A, B, etc.
ED-5.05 Definitions
ED-5.06 Evidence.
ED-5.07 Plausible Range.
ED-5.08 Individual Elicitation .
ED-5.09 Fitting (SHELF Stage-1).
ED-5.10 Group Discussion (SHELF Stage-2 begins).
ED-5.11 Group Plausible Range.
ED-5.12 Group Elicitation.
ED-5.13 Fitting/Feedback (SHELF Stage-2 ends).
ED-5.14 Chosen Distribution (SHELF Stage-3).
ED-5.15 Discussion.
ED-5.16 End Time of Workshop Part 2 (Elicitation).

For developing the SHELF exercise example, we choose to work with the example of a UQ problem for the length of a crack described in **Approach-2 (DEX).** Instead of asking 8 teams to

examine a weldment for the length of a subsurface crack, we choose to work with only 4 UT teams.

During workshop Part 2, the Facilitator will ask the 4 experts two sets of questions as shown in Fig. 10. The Recorder will display the fitting of the individual judgments of the four experts, A, B, C, and D, in the left of Fig. 11 under Stage-1.

A 4-step Stage-2 follows the display of individual distributions. The 4 steps are:

(1) Group Discussion. Each expert is requested by the Facilitator to state the basis of his or her judgment in arriving at the tertile-based subjective probabilities, and all experts and the Facilitator will contribute to a group discussion on the differences in the nature and quality of bases used by the experts.

(2) Group Plausible Range. The Facilitator will then initiate a discussion of a "group plausible limit" for each of the five limits available for discussion, namely, the lower bound, the upper bound, the lower tertile, the upper tertile, and the mean.

(3) Group Elicitation. At the end of Step (2), a consensus set of five numbers emerges as the group consensus judgment. One of the key features in using SHELF for expert knowledge elicitation is *not* to combine the four individual distributions into a single one by a statistical "averaging" method known as "linear pooling." The group consensus judgment is to be established after a face-to-face group discussion as outlined above in Step (1), and an agreement by all experts to a group plausible range as outlined in Step (2), and definitely *not* by the linear pooling method.

If a deadlock is reached during Step (2) for arriving at any single group consensus limit, the Facilitator will invoke another feature of SHELF for conflict resolution known as the postulated existence of a "hidden objective observer," whose task is to break the deadlock and deliver a group consensus through the Facilitator. This unique feature of SHELF allows the Facilitator to declare at the end of this step that a group consensus set of five numbers has been reached and recorded with the details of the disagreement that caused the deadlock.

(4) Fitting and Feedback. During this step, the Recorder will display a number of distributions that fit the group consensus judgment and ask for an agreement among all experts on the "best fit" that will serve as the key result of this knowledge elicitation exercise.

As shown on the right of Fig. 11, a consensus distribution is displayed during the Stage-3 of the SHELF workshop. The workshop ends with more discussion on the chosen distribution.

6. DISCUSSION

Of the three NDE UQ approaches presented in this paper, the most cost-effective and time-saving one appears to be **Approach-3 (EKE)** as shown in the following comparison table:

Table 7. A Comparison of 3 Approaches to NDE UQ

	Approach-1 (Regression Models)	**Approach-2** (Design of Experiments, 5 factors max.)	**Approach-3** (Expert Knowledge Elicitation)
1. Team Nos.	7 - 23	4 - 32	2 - 4
2. Time	Years	Days - Months	Days
3. Cost	> US$ 10^6	< US$ 10^6	< US$ $2x10^5$
4. Easy to update	***No.***	***No.***	Yes.
5. UQ Type	Global	Local	Local
6. Confidence Bounds Type	Predictive Intervals *(Not good enough###)*	Predictive Intervals *(Good for Local UQ)*	Predictive Intervals *(Good for Local UQ)*
7. POD UQ	Yes.	Yes.	Yes.
8. Length UQ	Yes.	Yes.	Yes.
9. Depth UQ	***No.***	Yes.	Yes.

###*Note-3: Regression models used in* **Approach-1** *were based on NDE data from a finite number of inspections, and the UQ results from their analyses were unfortunately limited to the plotting of the so-called "predictive intervals," which is valid only for predicting the result of the next inspection, and not that of any fraction of the entire population of future inspections. To rectify this deficiency, we need to introduce the concepts of "tolerance intervals" and "coverage," which are well covered in the statistics literature such as Refs. [32-37] and were applied by Fong, Hedden, Filliben, and Heckert [8] in 2008.*

However, NDE UQ results from **Approach-3 (EKE)** is local in nature, thus of little value to other NDE reliability users because the results will most likely be proprietary (unpublished) and certainly not generic enough. The next best cost-effective approach is **Approach-2 (DEX)**, and that is what NASA has done as recently reported by Generazio [21, 22]. Because **Approach-2 (DEX)** is evidence-based, we concur with NASA that it is the best among the three approaches outlined in this paper.

7. CONCLUDING REMARKS

Recent advances in sensor technology, data analysis and communication infrastructure, and web-based portals for field-office-field surveillance of NDE-monitored critical structures and components, have empowered engineers to improve their skills to manage aging infrastructure and prolong plant lives within a comfortable safety margin, provided we can quickly close the UQ gap in NDE reliability.

We have shown in this paper that closing the NDE UQ gap is technically feasible, but rather costly and time-consuming. The trade-off between the benefits of plant life extension plus public safety and the cost of a sizable investment in NDE reliability research is, in our opinion, very much in favor of the former, i.e., more NDE reliability research will prolong the service lives of key components of the society's civil infrastructure.

8. REFERENCES

[1] Adamonis, D. C., and Hughes, E. T., 1979, "Ultrasonic Evaluation and Sectioning of PVRC Plate Weld Specimen 201," *WRC Bulletin 252*, September 1979. New York, NY: Welding Research Council, 1979.

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

[3] Yukawa, S., 1986, "Sectioning Results for PVRC Plate Weld Specimen 251J," in *Proceedings of June 1983 Portland Symp. On NDE Reliability Through Round Robin Testing, ASME Spec. Pub. NDE-1*. New York, NY: ASME, 1986.

[4] Fong, J. T., 1986, "Analysis of Sectioning Data of PVRC 251J for Estimating Flaw Fabrication Reliability," in *Proceedings of June 1983 Portland Symp. On NDE Reliability Through Round Robin Testing, ASME Spec. Pub. NDE-1*, pp. 89-116. New York, NY: ASME, 1986.

[5] Fong, J. T., and Filliben, J. J., 1986, "A Data Analysis Methodology for Estimating NDE Reliability from Round Robin Ultrasonic Testing Data," in *Proceedings of June 1983 Portland Symp. On NDE Reliability Through Round Robin Testing, ASME Spec. Pub. NDE-1*, pp. 201-240. New York, NY: ASME, 1986.

[6] Doctor, S. R., 2008, "The History and Future of NDE in the Management of Nuclear Power Plant Materials Degrada-tion," in *Proc. ASME PVP Conf., July 27-31, 2008, Chicago, IL,* Paper PVP2008-61826. New York, NY: ASME, 2008.

[7] Bush, S. H., and Hedden, O. F., 2007, "Flaw Detection, Location, and Sizing," in *Proceedings Digest of ASME PVP Symp. On Engineering Safety, Applied Mechanics, and Non-destructive Evaluation (NDE) in honor of Dr. Spencer H. Bush (1920-2005),* July 2007, San Antonio, TX, pp. 153-156. Stanford, CA: Stanford Mechanics Alumni Club c/o Mechanical Engineering Department, Stanford University, 2007.

[8] Fong, J. T., Hedden, O. F., Filliben, J. J., and Heckert, N. A., 2008, "A Web-based Data Analysis Methodology for Estimating Reliability of Weld Flaw Detection, Location, and Sizing," in *Proceedings of 2008 ASME Pressure Vessels and Piping Division Conference*, July 27-31, 2008, Chicago, IL, U.S.A., Paper PVP2008-61612. New York, NY: ASME, 2008.

[9] Becker, F. L., 2001, "Examination Effectiveness Based on Performance Demonstration Results for the Flaws at the RPV Clad-to-Base-Metal Interface," in *Proceedings of Third International Conference on NDE in Relation to Structural Integrity for Nuclear Pressurized Components*, Nov. 14-16, 2001, Seville, Spain. Published in 2003 as *Report EUR 20671 EN*, pp. 690-698. Petten, The Netherlands: European Commission Joint Research Centre, Institute for Energy, 2003.

[10] Becker, F. L., 2002, "Reactor Pressure vessel Inspection Reliability," in Proceedings of EC-IAEA Technical Meeting on Improvements in In-Service Inspection Effectiveness, Nov. 19-21, 2002, Paper P14. Petten, The Netherlands: European Commission Joint Research Centre, Institute for Energy, 2003.

[11] EricksonKirk, M., Junge, M., Arcieri, W., Bass, B. R., Beaton, R., Bessette, D., Chang, T. H. J., Dickson, T., Fletcher, C. D., Kolaczkowski, A., Malik, S., Mintz, T., Pugh, C., Simonen, F., Siu, N., Whitehead, D., Williams, P., Woods, R., and Yin, S., 2007, *Technical Basis for Revision of the Pressurized Thermal Shock (PTS) Screening Limit in the PTS Rule (10 CFR 50.61), NUREG-1806,* Vol. I. Washington, DC: U.S. Nuclear Regulatory Commission, 2007.

[12] Gosselin, S. R., and Simonen, F. A., 2009, "Accounting for the Effects of Inservice Inspection Flaw Depth Sizing Errors in Developing Flaw Size Distribution Tables," in Pro-ceedings of 2009 ASME Pressure Vessels and Piping Div. Conf., July 26-30, 2009, Prague, Czech Republic, Paper PVP2009-77559. New York, NY: Amer. Society of Mechanical Engineers,, 2009.

[13] Heasler, P. G., and Doctor, S. R., 1996, "Piping Inspection Round Robin," *NUREG/CR-5068, PNNL-10475.* Washington, DC 20555: U. S. Nuclear Regulatory Commission, 1990.

[14] Heasler, P. G., Taylor, T. T., Spanner, J. C., Doctor, S. R., and Deffenbaugh, J. D., 1990, "Ultrasonic Inspection Reliability for Intergranular Stress Corrosion Cracks - A Round Robin Study of the Effects of Personnel, Procedures, Equipment and Crack Characteristics," *Nuclear Regulatory Commission Report, NUREG/CR-4908, PNL-6196.* Washington, DC 20555: U. S. Nuclear Regulatory Commission, 1990.

[15] Lemaitre, P., and Borloo, E., 1994, "Non-Destructive Examination Practice and Results, State of the Art and PISC III Results," *Proceedings of the Joint CEC OECD IAEA Specialists Meeting held at Petten*, Joint Research Centre, Institute for Advanced Materials, Box 2, NL-1755, ZG Petten, The Netherlands, ISBN 92-826-8813-5, 1994.

[16] Seuaciuc-Osorio, T., and Ammirato, F., 2017, "Materials Reliability Program: Development of Probability of Detection Curves for Ultrasonic Examination of Dissimilar Metal Welds (MRP-262, Revision 3): Typical PWR Leak-Before-Break Line Locations," *EPRI 2017 Technical Report No. 3002010988*. Palo Alto, CA 94304: Electric Power Research Institute, 2017.

[17] Heasler, P. G., and Doctor, S. R., 2003, "A Comparison of Three Round Robin Studies on ISI Reliability of Wrought Stainless Steel Piping," *Nuclear Regulatory Commission Report,*

NUREG/CR-6795, PNNL-13873. Washington, DC 20555: U. S. Nuclear Regulatory Commission, 2003.
[18] 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.
[19] John, P. W. M., 1990, *Statistical Methods for Engineering and Quality Assurance*. Wiley, 1990.
[20] Wu, C. F. J., and Hamada, M., 2000, Experiments: Planning, Analysis, and Parameter Design Optimization. Wiley-Interscience, 2000.
[21] Generazio, E. R., 2009, "Design of Experiments for Validating Probability of Detection Cpability of NDT Systems and for Qualification of Inspector," Materials Evaluation, Vol. 67, No. 6, pp. 730-738, 2009.
[22] Generazio, E. R., 2015, "Directed Design of Experiments for Validating Probability of Detection Capability of NDE Systems (DOEPOD)," *NASA Report, NASA/TM-2015-218696*, 2015.
[23] Winkler, R. L., 1967, "The quantification of judgment: Some methodological suggestions," J. Am. Stat. Assoc., Vol. 62, pp. 1105-1120, 1967.
[24] Murphy, A. H., and Winkler, R. L., 1977, "Reliability of subjective probability forecasts of precipitation and temperature," *Appl. Statist.,* Vol. 26, pp. 41-47, 1977.
[25] Cooke, R., M, 1991, *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford: Oxford Univ. Press.
[26] O'Hagen, A., Glennie, E. B., and Beardsall, R. E., 1992, "Subjective modelling and Bayes linear estimation in the UK water industry," *Appl. Statist*., Vol. 41, pp. 563-577, 1992.
[27] O'Hagan, A., 1998, "Eliciting expert beliefs in substantial practical applications," *The Statistician*, Vol. 47. Part 1, pp. 21-35. 1998.
[28] O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T., 2006, *Uncertain Judgements: Eliciting Experts' Probabi-lities*. Wiley, 2006.
[29] Oakley, J. E., and O'Hagan, A., 2007, "Uncertainty in prior elicitations: a nonparametric approach," *Biometrika*, Vol. 94, No. 2, pp. 427-441, 2007.
[30] Gosling, J. P., 2018, "SHELF: The Sheffield Elicitation Framework," Chapter 4 in *Elicitation: The Science and Art of Structuring Judgement,* Dias, L. C., Morton, A., and Quigley, J., Editors. Springer, 2018
[31] O'Hagan, A., 2018, *Elicitation with SHELF*, a Course book, January 8-9, 2018, tony@tonyohagan.co.uk, 2018.
[32] Natrella, M., 1963, *Experimental Statistics*, National Bureau of Standards Handbook No. 91. Washington, DC: U.S. Government Printing Office, 1963.
[33] Mandel, J., 1964, The Statistical Analysis of Experimental Data. New York, NY: Interscience, 1964. Reprinted by Dover, 1984.
[34] Guttman, I., 1970, *Statistical Tolerance Regions: Classical and Bayesian*. London: Charles Griffin & Co., Ltd, 1970.
[35] Hahn, G. J., and Meeker, W. Q., 1991, *Statistical Intervals: A Guide for Practitioners*. Wiley, 1991.
[36] Nelson, P. R., Coffin, M., and Copeland, K. A. F., 2003, *Introductory Statistics for Engineering Experimentation*. Elsevier Academic Press, 2003.
[37] Meeker, W. Q., Hahn, G. J., and Escobar, L. A., *Statistical Intervals: Guide for Practitioners & Researchers*. Wiley, 2017.

A UT Field Work Example

(Proprietary data changed to protect owner's IP)

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

Fig. 1. A 2-level, 5-factor experimental program for an ultrasonic testing (UT) inspection of a crack length problem [8].

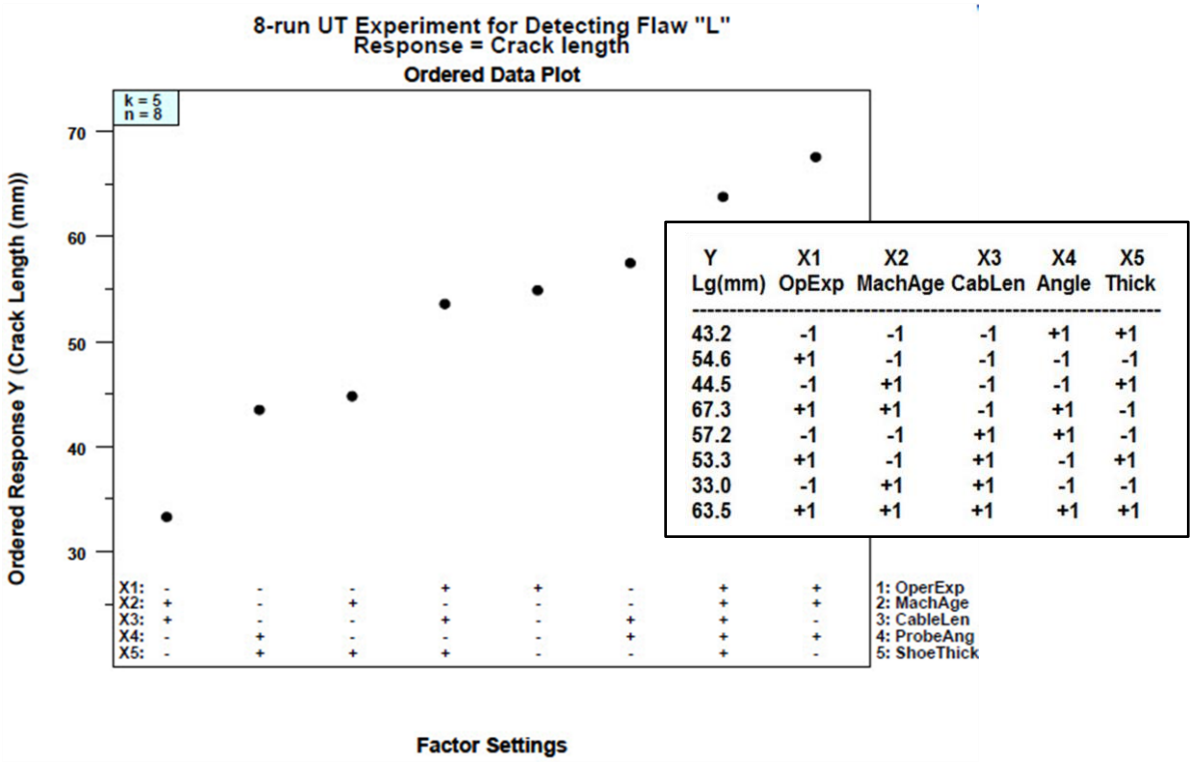

Fig. 2. A 2-level, 5-factor, 8-run fractional factorial design of a UT experiment for finding crack length (in.) [8].

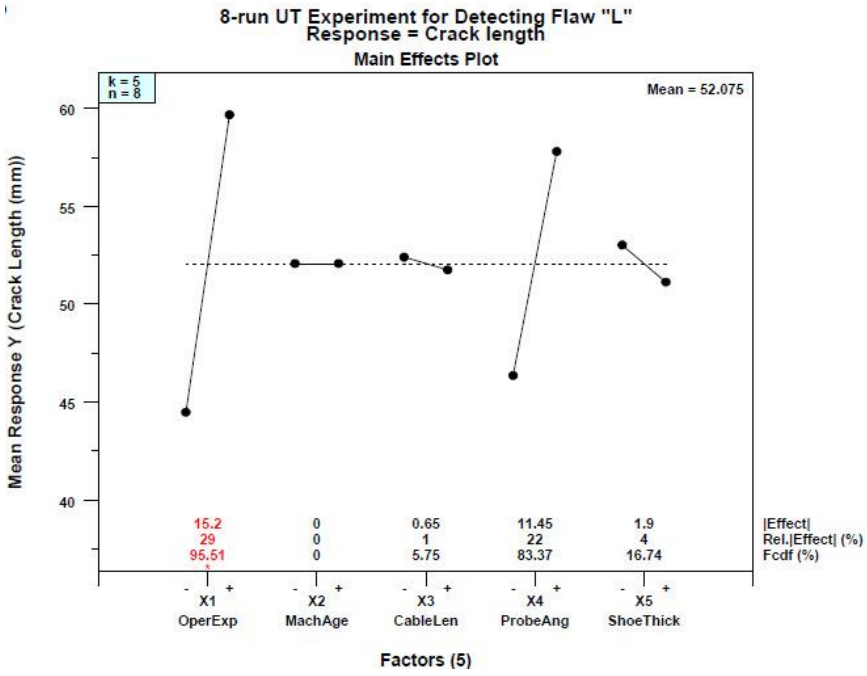

Fig. 3. Plot of Main effects from the analysis of a 2-level, 5-factor, 8-run Design of Experiments (DEX) [8].

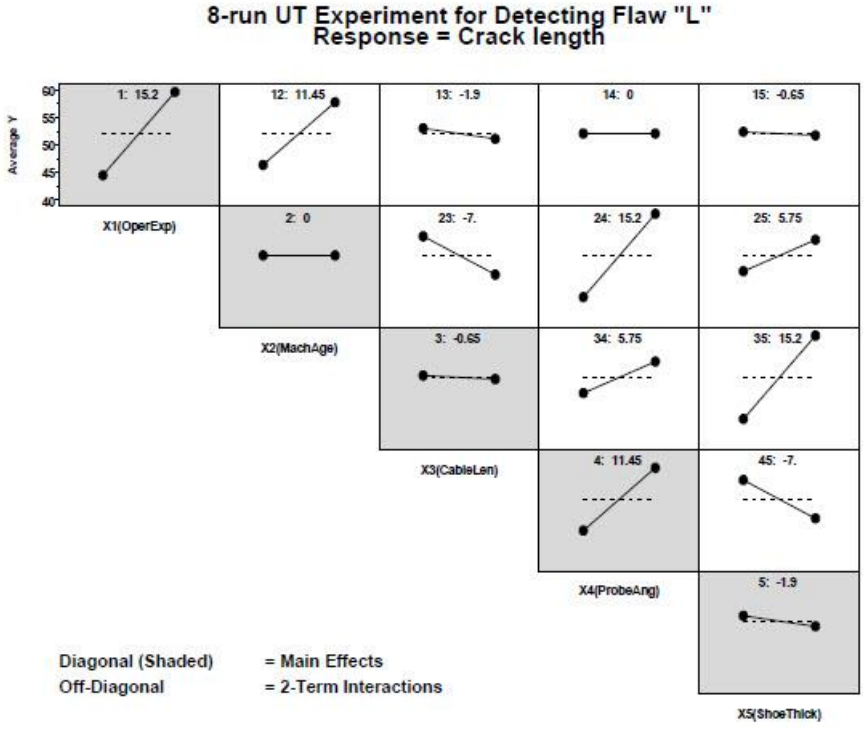

Fig. 4. Interaction effects matrix from the analysis of a 2-level, 5-factor, 8-run DEX [8].

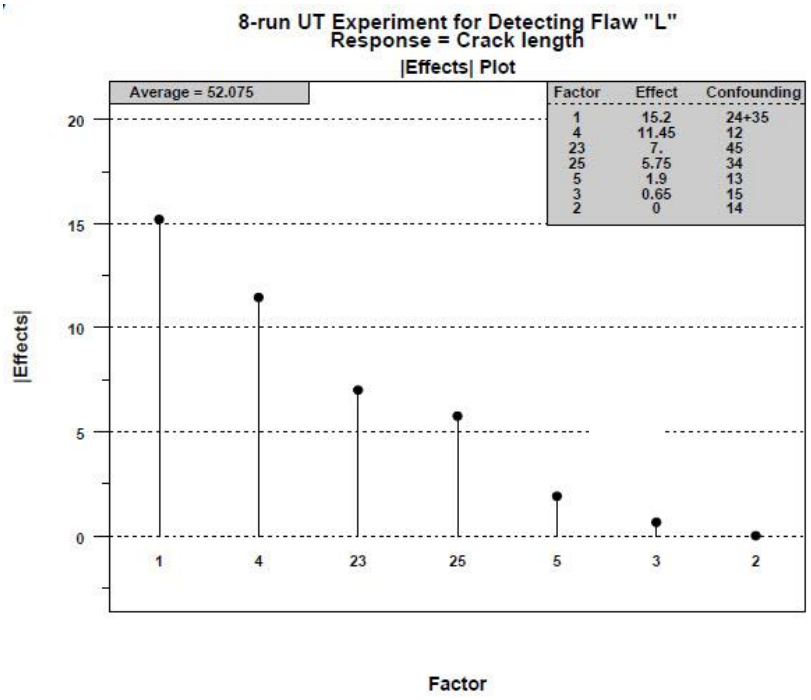

Fig. 5. | Effects| plot from the analysis of a 2-level, 5-factor, 8-run Design of Experiments (DEX) [8].

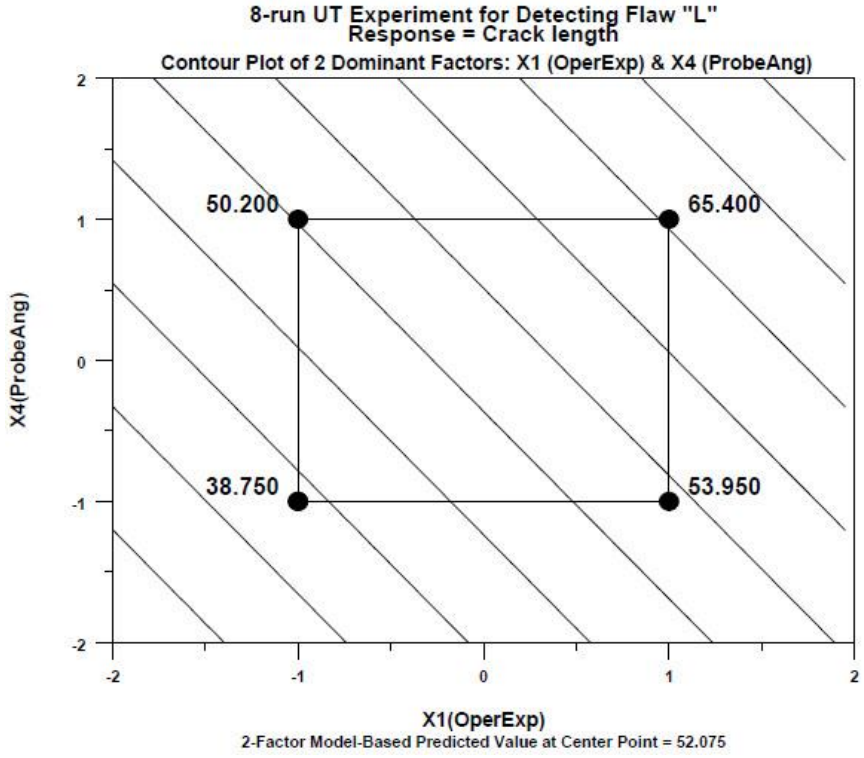

Fig. 6. Contour Plot of 2 dominant factors of a 2-level, 5-factor, 8-run DEX [8].

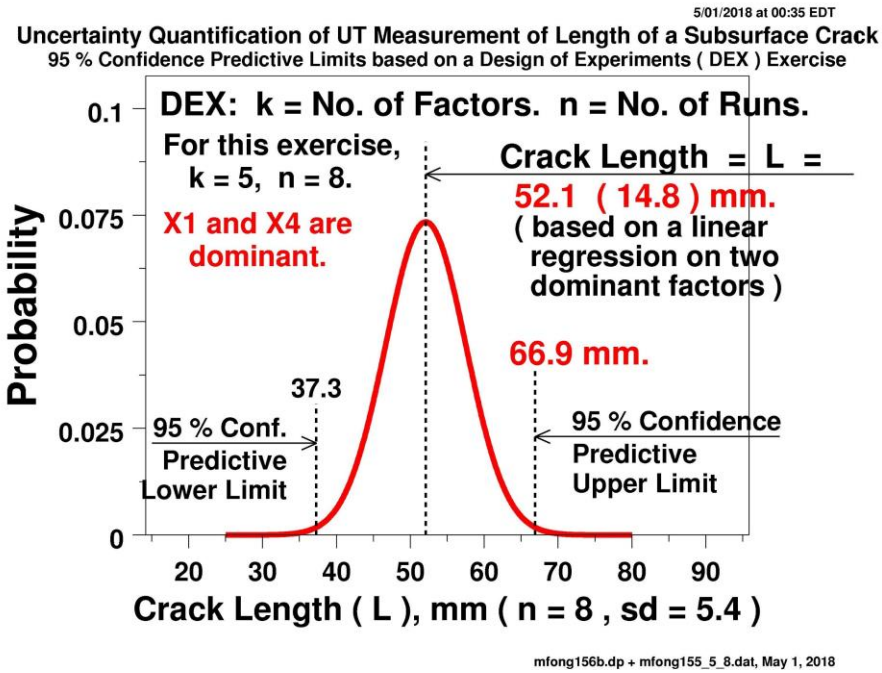

Fig. 7. Crack length (mm.) distribution plot from the analysis of a 2-level, 5-factor, 8-run DEX [8].

Fig. 8. A pre-elicitation flow chart of a typical SHELF exercise with eight distinct activity boxes (after [31]).

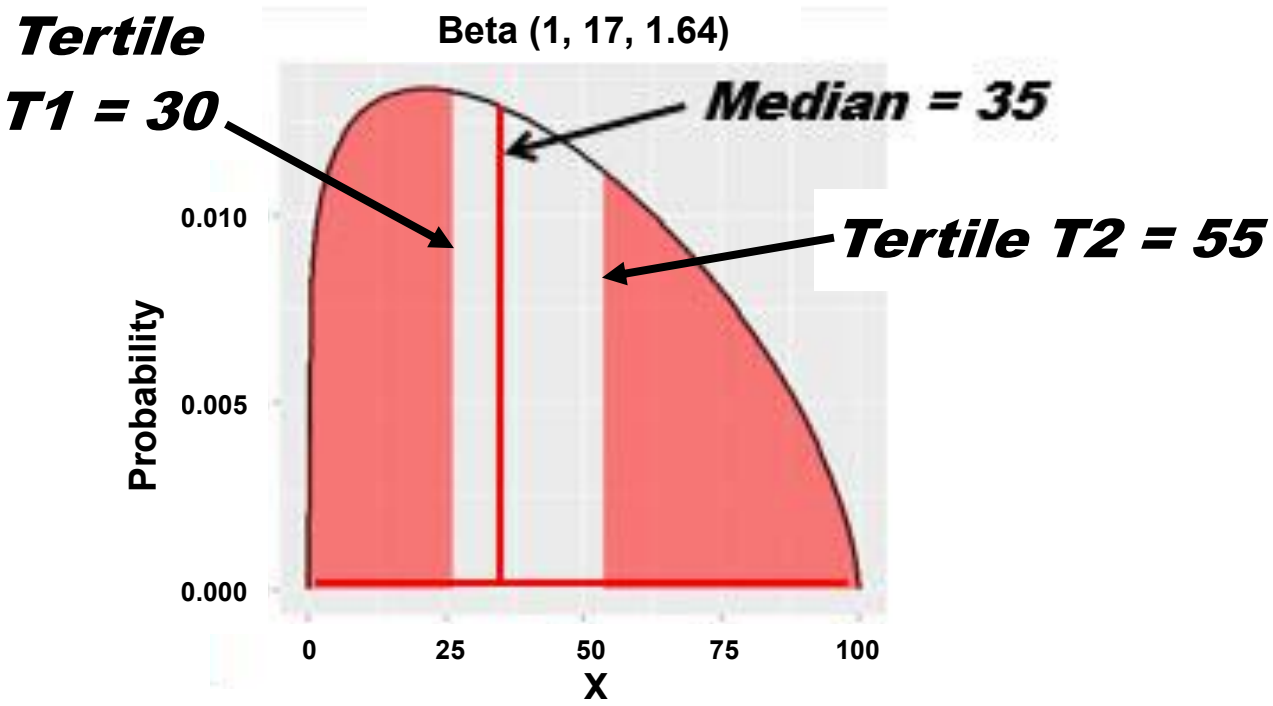

Fig. 9. A parameter fit of a typical response of an expert when asked to estimate an unknown quantity in terms of five subjective probabilities, namely, 0 % for a lower bound (L = 0), 33 % for a lower tertile (T1 = 30), 50 % for a median (M = 35), 67 % for an upper tertile (T2 = 55), and an upper bound (U = 100).

Two sets of questions were asked:

Set-1: Name the lower bound, **L** , median, **M** , and upper bound, **U** , of the *QoI*.

Set-2: Name the lower **tertile**, **T1** , and the upper **tertile**, **T2** , of the *QoI*.

Expert	Lower Bound, L	Lower Tertile, T1	Median M	Upper Tertile, T2	Upper Bound, U
A	0.90	*1.55*	1.70	*1.85*	2.70
B	1.10	*1.95*	2.15	*2.40*	3.10
C	1.00	*1.60*	1.75	*1.95*	2.80
D	2.00	*2.50*	2.65	*2.85*	3.50

Fig. 10. Estimates of crack lengths by 4 experts as a function of 5 subjective probabilities in a SHELF exercise example [31].

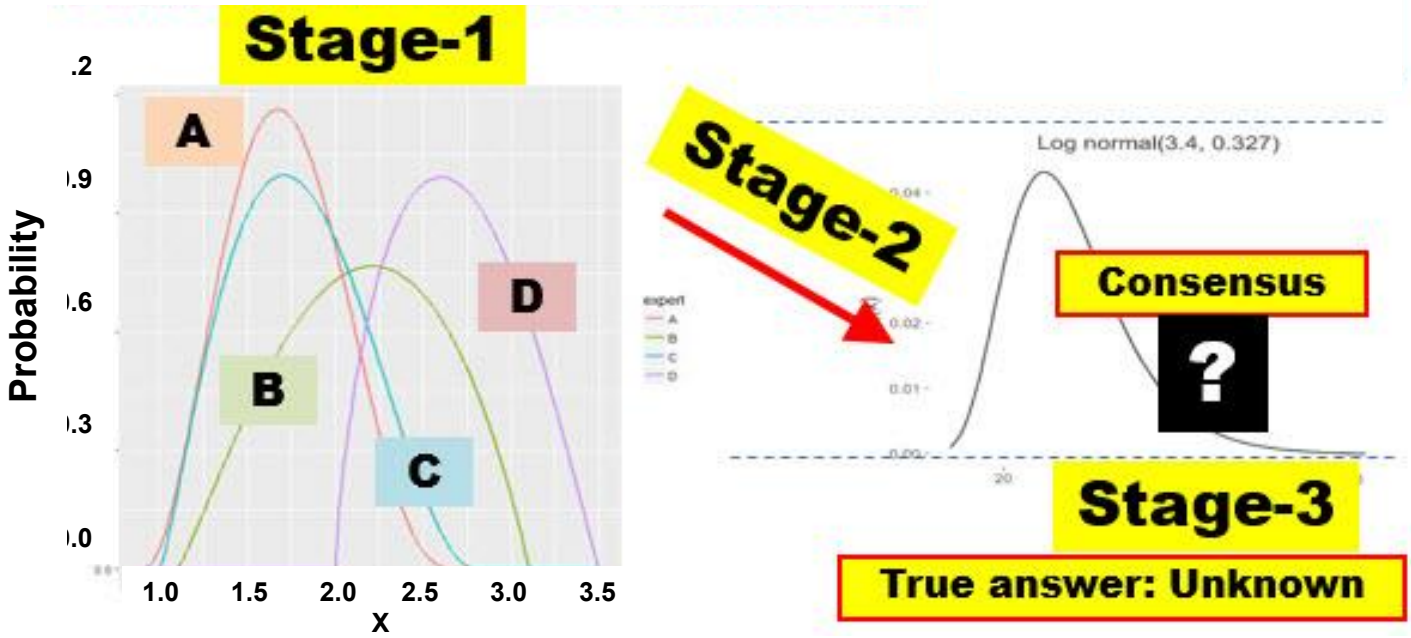

Fig. 11. The 3-stage SHELF exercise example for finding the crack length of a weldment with confidence bounds [31].