Holistically Evaluating the Reliability of NDE Systems – Paradigm Shift

Christina MUELLER 1, Marija BERTOVIC 1, Mato PAVLOVIC 1, Daniel KANZLER 1, Uwe EWERT 1, Jorma PITKÄNEN 2, Ulf RONNETEG 3

1 Federal Institute for Materials Research and Testing, Unter den Eichen 87, 12205 Berlin, Germany, Phone: +49 30 81041833, Fax +49 30 81041836; e-mail: Christina.mueller@bam.de)
2 Posiva Oy, Eurajoki, Finland
3 SKB, Oskarshamn, Sweden

Abstract
New methodologies for evaluating the reliability of NDE systems are discussed in accordance with the specific requirements of industrial application. After a review of the substantive issues from the previous decades, the go forward guidance is concluded.

For high safety demands a quantitative probability of detection (POD) created from hit miss experiments or signal response analysis and ROC (Receiver Operating Characteristics) are typically created. The modular model distinguishes between the influence of pure physics and technique, industrial application factors and the human factor and helps to learn what factors are covered by modelling, open or blind trials. A new paradigm is offered to consider the POD or reliability of the system as a function of the configuration of input variables and use it for optimisation rather than for a final judgement. New approaches are considered dealing with real defects in a realistic environment, affordable but precisely like the Bayesian approach or model assisted methods.

Among the influencing parameters, the human factor is of high importance. A systematic psychological approach helps to find out where the bottlenecks are and shows possibilities for improvement.

Keywords:
Reliability of NDE, POD, ROC, Modular Model, Multiparameter-POD, Bayesian Approach, Human Factor

1. Review of Approaches to NDE Reliability

1.1. State of the Art as described by the European-American Workshops on NDE Reliability 1997 – 2002

Before presenting the new paradigm in the field of NDE reliability the authors would like to summarize the insights from the first three workshops from the series of European-American Workshops on NDE reliability. The main achievement from the first workshop 1997 in Berlin was the conceptual model in terms of the reliability formula which simply says that the total reliability of an NDE system is composed of the intrinsic capability, IC (which stands for the physical principle behind the defect indication and its technical realization as an upper bound), the application factors, AP (describing the realistic circumstances like UT-coupling, limited access, noise of the surrounding …) and the human factor, HF, present in each application. While imperfect – e.g. the mutual interactions between the factors were not considered – the conceptual model helped to properly define potential for performance optimization, and also as an assessment tool for the adequacy of open and blind trials. A main
benefit from the Second European American Workshop on NDE Reliability, September 1999, Boulder, Colorado, USA [1], was the clear definitions of i) the **NDE System** as the procedure, equipment and personnel that are used in performing NDE inspection and ii) the **NDE reliability** as the degree that an NDT system is capable of achieving its purpose regarding detection, characterization and false calls. The main conclusion 2002 from the third European American Workshop was: We need to **quantify the risk in NDE and demining!**

From the beginning, it was clear that we are asking the NDE process to describe the real status of a component. Later, limits were discovered for NDE systems e.g. for small flaws, not all flaws will be detected. The second task then is to provide quantitative reliability information about inspection results at least for defined risk critical applications. In Figure 1 this means that there is not always a 1:1 correspondence between the “truth” (the real defect situation) in a weld (left side) and the defects indicated by the NDE system which are recorded in the inspection record (right side).

The traditional European approach relies on standards i.e. to describe value ranges and allowed uncertainties of the physical and technical parameters in a standard and test according to the standard as illustrated in Figure 2.
An alternative approach – the performance demonstration strategy as required by ASME section XI appendix VIII – employed by ISI NDE for the US nuclear industry represented by EPRI since the 1980s – is illustrated in Figure 3. The NDE system – composed of the NDE equipment, the procedure and the human personnel is handled as a black box. Only the “input” in terms of the flawed test samples is compared to the “output” the test results. It is the responsibility of the NDE provider how to reach the goal and document the practice in a procedure/specification. The results are then statistically evaluated and acceptance limits are set.

Figure 4 illustrates the ROC (Receiver Operating Characteristics)-curves as a performance measure which has been used for a number of decades in the operation of radar, and also in medical diagnostics [2, 3]. Within the ROC-scheme for the characterization of the performance of an NDE system the proportion of hits (true positive indications) are evaluated against false calls (false positive indications) in a diagram. Along one ROC-curve the sensitivity is rising as the decision threshold is lowered, i.e. since the threshold aims to distinguish between noise and wanted signals, more and more noise signals are accumulated. In the upper right corner all detected signals are gathered. The quality of NDE-systems raises from the straight line – the “pure chance” curve where noise and wanted signals are counted with the same frequency - up to the left edge step curve where only valid signals are gathered (ideal system). The truth of real NDE systems might occur on all ranges between theses extremes (see e.g. Figure 15). For NDE-engineers it is now important to learn what physical parameters do influence the probability of detection to be limited or growing. For a fixed false call rate – as indicated in Figure 5 – we look for the value of the POD and how it is distributed. Originating in the US-aerospace industry [4] another data sampling scheme is utilized for this purpose – the “Probability of Detection” (POD) as a function of defect size (or another appropriate defect parameter) with Figure 6 showing the result for one applied threshold.
For automated thresholding systems – e.g. eddy current testing of gas turbine engines – the method has been developed further for a signal response POD as illustrated in Figure 7.
A defect of size “a” causes a signal of height “â”. When the signal is above the selected threshold it is counted as defect indication and otherwise as noise. The “â versus a” curve – in almost all cases on a logarithmic scale is indicated on the lower left hand side of Figure 7. Using an appropriate statistical model for the data distribution (see [4]) this diagram can be transformed into a typical POD curve with a corresponding 95% lower confidence bound.

1.2. New Views and Approaches 2009-2012: New methodologies in evaluating the reliability of NDE systems unerring, reliable and efficiently according to the specific requirements of industrial application

In the above chapter we discussed the issues already known during the 3rd European-American Reliability workshop in 2002. What has developed since?
With the “NDE Reliability Flower Garden” shown in Figure 8 the authors like to illustrate the considerable and versatile progress all over the world between the corner points of industrial demands and scientific progress on the one hand and between safety and security demands and economic conditions on the other hand. What is the blooming item in each of the flowers? As indicated in the work from EPRI [5] the PD (Performance Demonstration) is now well established in the US nuclear industry and with its procedure examination it drew nearer to the ENIQ (European Network of Inspection Qualification) approach, which was from the beginning composed of parameter assessment and open and blind trials. The National adaptations of the ENIQ methodology are also spread out over Western and Eastern Europe today [6]. The POD-qualification of NDE methods applied to airplane components is mandatory worldwide. For general NDE applied by accredited labs (EN-ISO 17025) the validation of quantitative and qualitative testing methods is required when no established standard exists. Also the new branch of SHM (Structural Health Monitoring) is going to evaluate the reliability via POD curves. RBI (Risk Based Inspection) and RBLM (Risk Based Life Time Management) is more and more applied to pressure equipment and nuclear facilities and the reliability of NDE is an important bridge between testing and the assessment of the remaining risk and life time. The youngest most promising flower is the modelling assisted POD (MAPOD see Bruce Thompson, et. al. [7]) which helps to reduce costs for a high number of empirical experiments by substituting missing information by modelled results. Last not least for dealing with the Human Factor in an appropriate scientific way working psychological methods came on board firstly touched in investigations within the PANI program and investigations in demining and NDE by BAM. But even more insights from this field are to be taken into account when completing the modular reliability model: the organizational context within which all the modules are embedded and also the interaction between different organizations as we can learn from Babette Fahlbruch [12]. The generalized modular model as depicted in Figure 9 can also help to understand what issues are covered by the different “Flowers” of Figure 8.
Generally, when asked for an assessment of the reliability of an NDE-system the following approach is helpful to consider: first, the actual level of safety demands have to be defined in order to adapt the thoroughness of investigation to the level of risk when the tested component would fail. Then, all the essential influencing parameters need to be listed and transferred to an appropriate design of experiments to determine the reliability in terms of a qualitative assessment for lowering the risk, or in terms of quantitative probability of detection (POD) or ROC (Receiver Operating Characteristics) curves.

We recommend here a new paradigm which considers the POD or reliability of the system as a matrix of input variables utilized for process optimization, rather than for a final judgment. A subsequent element, and we believe advantage of this approach for the end user, is also to sample all single PODs to an integral “Volume POD” of a part including data fusion [9].

Among the influencing parameters, the human factor is the most important one but of course after the basics via IC are established. A systematic psychological approach should help to find out where the bottlenecks are, but as first priority, to provide the best possible working conditions for the human inspectors.

![Modular Model for NDE-Reliability](image)

Figure 9: The modular reliability model helps to understand and weight the different influences as realized by MAPOD and ENIQ, PANI and BAM

2. Application Examples for the Progress in Methods for Determining Intrinsic Capability, Application Parameters and Human Factors

2.1. Examples for new handling of Intrinsic Capability and Application Parameters

Multiparameter POD
Figure 10 shows a typical result using the “â versus a” scheme as explained above. When applying this approach to industrial applications like copper canister components for radioactive waste (see e.g. [8], [9], [10], [14]) we have to expand this approach to real industrial conditions.

Specifically, we see a need for a multi-parameter “a“ (depth, size, orientation, roughness …) and a data-field „â“ (more than a maximum signal) for industrial applications with thick parts and complex defect shapes.

To investigate this a modelling-assisted multi-parameter methodology (Pavlovic [8], [9]) was developed and applied to the lid of the copper canister seen in Figure 11. The POD as a function of defect size (FBH diameter), defect depth and defect angle is presented in Figure 12(a), (b) and (c), respectively. The sharp decline of POD with increasing angle, for example, shows how important a comprehensive multi-parameter consideration is.

![Figure 10: Probability of Detection (POD) – â vs. a approach](image)

![Figure 11: Copper Canisters for deep deposit SKB/Posiva, Canister Parts planned for the NDT inspection: Copper Lid](image)
The investigation examples presented so far include the consideration of the intrinsic capability (IC) and also parts of the application factors (AP).

**Bayesian Approach for dealing with limited data from real defects**

Applying the POD evaluation to real industrial processes, the need to deal with the real occurring flaws under production or in service conditions arises as a natural application factor. To create those in a statistically sufficient amount is too expensive or even unfeasible. The Bayesian Approach provides a good solution to compute POD-curves in case of small amount of real defects without losing the necessary information by incorporating data from artificial defects via a prior function. The posterior function contains the needed information for the computation of POD-curves for real defects with an acceptable and sufficient amount of information, even for sparse amount of data as shown in the work of Kanzler, et. al. [10].
Figure 14: The POD made of only real defects and the improvement with the Bayesian approach with combining real and artificial data

Summary result of a round robin test in radiographic weld inspection
The next example – as contrast to the almost “pure” examples covering intrinsic probability and application factors - shows ROC curve results in Figure 15 for the same technique (X-ray inspection) applied to weld specimens evaluated by 20 different inspectors. So, the difference in the results is only caused by the difference in human performance (see [11]).

Figure 15: ROC of 20 human inspectors for radiographic film evaluation from [5]

- Radiographic film images of 40 weld sections were evaluated by 20 inspectors
- grey: the ROC-curves of the individual inspectors
- red: the overall mean-value-curve with the actual operating points and error bars in the maximum point

The high diversity in this and other similar investigations suggests that we need to understand human factors in a more systematic way making advantage of the facilities of working psychology.
2.2. Human Factors

According to the periods of safety research (see Figure 16), developed by Reason (1993; in [12] and expanded by Wilpert & Fahlbruch (1998; in [12]), the different influencing factors on the reliability of safety relevant work results – from technology over individual human errors to socio-technical and inter-organizational factors - are well known for a while in the human factors field. Within the human factor project SR2514, sponsored by the German’s Federal Office for Radiation Protection (Bundesamt für Strahlenschutz, BfS), on human factors’ influences on manual ultrasonic inspection performance during ISI in nuclear power plants, a human factor model had been set up and tested which contained the three first areas of factors of Figure 16 (see [13] for detailed results).

Figure 16: Periods of Safety Research

The specific goal of the project was to determine if and how time pressure influences the manual ultrasonic inspection performance. Based on a DoE protocol, ten experienced inspectors carried out manual ultrasonic inspection. Additionally two teams carried out mechanised/automated inspection. All inspections were accomplished on a mock-up reactor vessel and on various test pieces with programmed flaws (notches and cracks). The selected flaws were typical for tasks performed during ultrasonic in-service inspection in nuclear power plants. The task was to find the flaws in given areas and to determine the corresponding amplitudes, coordinates and length dimensions. The evaluation criteria for the quality of inspection and for the human factors’ influence, was the measurement precision for detected indications.

To substantiate the human factors element, the main focus was put on parameterization of the manual inspections. The time available for inspection, evaluation and writing the protocols was varied in three levels (A-no time limitation, B-middle time pressure, C-high time pressure). Psychological factors, i.e. mental workload, stress resistance and stress reaction were additionally assessed in the course of the experiment.

The results show a high influence of human factors on the inspection results. The main hypothesis was confirmed, i.e. that the time pressure has an effect on the measurement precision. The effect of individually perceived time pressure (see Figure 17) (rather than the so called “objective” setting from outside) and mental workload were found to be significant having a negative impact on the performance. A good preparation before the inspection (e.g.
thorough briefing and training on test pieces) and the communication between the "organisation" (e.g. the supervision of the utilities, inspection companies and expert organisations) and the inspectors were considered as highly important. The influence of the organisation has shown to be relevant for the practice. When the inspectors consider the inspection tasks as not precise enough defined, the documentation as ambiguous and not manageable, the working conditions as inadequate, and above all when they feel they have not been sufficiently prepared for the tasks, there is a clear increase in the mental workload of the task and consequently a decrease in the measurement precision. Figure 18 shows differences between different operators in the amplitude estimation through all experimental conditions (time pressure A, B and C).

Figure 17: Results: the scatter of amplitude, extension and depth measurement as a function of the perceived temporal demand (A – Low temporal demand, B – Middle temporal demand, C – high temporal demand)

Figure 18: Measured Amplitude-mean values - Differences between the inspectors

Conclusions from the project:
- It was confirmed time pressure has an effect on the quality of UT inspections
- The organizational context determines the way inspections are performed and therefore highly influences on the inspection quality.

For an improvement of the human performance attention needs to be drawn on:
- Demonstration task for training and confirmation
• Organization – good preparation
• Written procedures and protocols
• Supervision

Further attempts to investigate human factors in the NDT-field have focused on the mechanized inspection of the canister, to be used for the final disposal of nuclear waste in Sweden and Finland. A substantial number of factors that could affect the reliability of NDT methods have been identified and analyzed. Proposals for the compensation of varying human performance, according to the experts, include the implementation of human redundancy (known also as the 4-eyes principle) or the semi-automation of the defect detection process. However, implementing human redundancy in critical tasks, such as defect identification, as well as using an automated aid (software) to help operators in decision making about the existence and size of defects, could lead to other kinds of problems, namely social loafing (exercising less effort when working on tasks collectively as compared to working alone) and automation bias (uncritical reliance on the proper function of an automated system without recognizing its limitations and the possibilities of automation failure) that might affect the reliability of NDT in an undesired manner, when not taken into account adequately (as elaborated by Bertovic, et. al. [14]). Deeper understanding of the defect detection and sizing process, as well as an evaluation of the existing procedures, are a part of the continuing effort to understand the varying human performance, as well as to optimize current practices, if needed.

3. Conclusion and Outlook

Dealing with the NDE reliability measurement effort the safety and economic requirements have to be taken into account and the qualitative and quantitative importance of each subtask (detection, false calls, characterisation, sizing) needs to be assessed. Multi-parameter POD and methods employing modelling assisted POD or data combination by Bayesian Approach appear promising to simultaneously fulfil safety and economic demands. The modular model helps to understand the impact of different sources of influences from parameters connected to physics, application or human factors – individual or organizational, respectively. Human factors need to be more deeply investigated by means of working psychology as ways to optimize the working conditions and preparative actions for the operators.

4. Acknowledgement

The authors would like to thank Lloyd Schaefer as well as Babette Fahlbruch for helpful discussions.

5. REFERENCES


9. Mato Pavlovic et. al., Th.3.A.2; Multi-Parameter Influence on the Response of the Flaw to the Phased Array Ultrasonic NDT System. The Volume POD; CD of Conference Proceedings of the 4th European-American Workshop on Reliability of NDE, June 24-26, 2009, DGZIP


