Effect of Feedback and Variation on Inspection Reliability

Iikka VIRKKUNEN [1], Jonne HAAPALAINEN [2], Suvi PAPULA [3], Teemu SARIKKA [3], Juha KOTAMIES [4], Hannu HÄNNINEN[3]
[1] Trueflaw Ltd., Espoo, Finland
[2] VTT, Espoo, Finland
[3] Aalto University, Espoo, Finland
[4] Metropolia University of Applied Sciences, Helsinki, Finland
Contact e-mail: iikka@trueflaw.com

Abstract. The reliability of Non-destructive testing (NDT) is an on-going challenge. The consequences of failed inspections can be dire, and thus the requirements for NDT reliability are very high. The work is technically demanding and requires skilled use of the available equipment and keen judgement to properly discern flaw signals from noise. Somewhat paradoxically, the work is also very tedious and repetitive. Most of the inspected targets do not contain any flaws but the inspectors need to be constantly alert for the possibility. The recent studies on human factors have brought advances in (among other things) improved readability of inspection procedures and procedures of reviews and redundant inspections widely used in order to improve overall inspection reliability.

In present paper, the effect of feedback and variation on inspector performance is studied. To test this, a small empirical study was completed. An online tool was created with a simplified UT set-up: B-scan image of the data and software gain control and tools to indicate cracks with point and click. The system generates random flawed data images on the fly. The user then analyses the images and indicates found flaws by clicking them. After 150 images have been analysed (many of them without flaws), the system uses the provided hits and misses to compute a POD curve and confidence bounds using standard (ASTM E2862) techniques. Additional "learning" version of the tool was created. In this "learning mode", after user requests next image, the system shows results of the current images (i.e. hits, misses and false calls in the current image). This set-up provides the inspector with direct feedback of his success and better facilitates learning this particular inspection task. The tool was presented to a small group of 9 inspectors in level-III inspector training and results were gathered from trainees both before and after training (and with and without feedback).

The results from this small group of inspectors indicated, that the direct feedback on achieved reliability can quickly improve POD values. However, the study group was small and thus the results need further investigation and confirmation.

Introduction

The reliability of Non-destructive testing (NDT) is an on-going challenge. The consequences of failed inspections can be dire, and thus the requirements for NDT reliability are very high. The work is technically demanding and requires skilled use of the available equipment and

keen judgement to properly discern flaw signals from noise. At the same time, the work is also very tedious and repetitive. Most of the inspected targets do not contain any flaws but the inspectors need to be constantly alert for the possibility. Finally, there are variation and challenges in the working conditions of the actual inspection: the inspections are often completed in awkward positions, uncomfortable high temperatures and the inspection target may not offer sufficient coupling with the inspection instruments.

Significant effort has been put into securing the performance and reliability of the inspections. Firstly, the inspections are codified into detailed procedures, for the inspector to follow. These written procedures enhance the repeatability and consistency over different inspectors and conditions. Numerous standards now exist for basic inspections [e.g. 1,2]. In addition, more detailed (and comprehensive) procedures for especially demanding or critical inspections are used, e.g. in the aerospace industry and the nuclear industry [e.g. 3]. The use of common standard practices and procedures greatly improves both the reliability and predictability of the inspection. Furthermore, it provides a common understanding of generic inspection capabilities and expected performance. Fixed procedures are also necessary precondition for meaningful measurement of expected performance.

However, as has been shown in various round-robin exercises [4], a common procedure is not necessarily sufficient to guarantee the needed performance. In the aerospace industry, rather conservative "default" performance levels have been adopted, that may be applied in absence of additional evidence [5]. In addition, methodology for quantitative assessment of actual performance has been developed [6]. In the nuclear industry, NDT qualification and performance demonstration were developed and are now required around the world for nuclear inspections [3,7]. The central idea in these is, that the performance of the procedure and then the performance of individual inspectors applying the procedure are verified with combination of technical justification and practical trials, where the inspection procedure is applied on test samples with known flaws and its performance is evaluated.

The performance demonstrations have significantly improved the expected inspection results. However, some high-profile failures in qualification inspection [8] have shown, that the actual inspection performance of a qualified inspector using procedure does not necessarily reflect the performance apparently demonstrated during qualification. These cases have led to increasing emphasis to "human factors" of inspection performance, i.e. factors affecting the inspection procedure but outside the traditional scope of performance of the technical inspection system and procedure [9, 10]. To account for the various separately studied aspects of inspection reliability, an integrated model has been proposed: "holistic approach to inspection reliability" [10], which offers the generic framework or synthesis for addressing various aspects of inspection reliability and their links and interdependencies. In the holistic approach, the inspection reliability is divided into four elements, as shown in Figure 1.

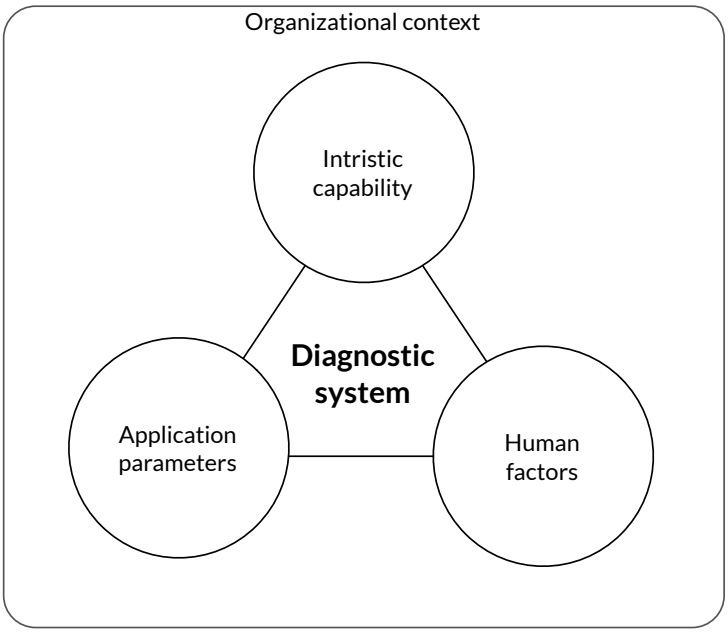

Fig. 1. Elements of inspection qualification according to the holistic approach for inspection [10].

The holistic approach has brought significant contribution to the field by making explicit the interconnectedness of various aspects of inspection reliability and the fallacy of estimating actual inspection performance based on demonstrated technical performance alone. However, the model, in its current form, has some shortcomings. Since the division to various components is done along empirical lines, all the components are deeply interconnected and addressing some aspect will potentially affect all the other components. This interconnectedness makes it difficult to use the approach for making quantitative estimates of inspection performance (even if all the components could be quantified in isolation, which at present is not possible, the combined performance is not apparent). Also, recent findings [9] indicate, that the inspectors do not necessarily learn or improve in terms of reliability (since the procedure is very detailed and kept constant, this is almost by design).

Furthermore, there are some evident contradictions inherent in the current approach. The procedures are made very detailed and adherence to the procedure is required of inspectors. The underlying assumption here is, that if the procedure fully defines the inspection and the inspectors fully follow the procedure, their performance is constant and predictable. In contradiction, it is well known that inspector performances vary significantly, and this is why personnel qualifications are required in addition to procedure qualifications in the nuclear industry. The procedures expect certain explicit (and implicit) preconditions to be met (e.g. surface quality of the inspection target, accessibility etc.). In practice, these are often violated, and the inspectors are put in contradictory position: they need to adhere to the procedure, which is impossible or counterproductive and at the same time obtain the performance attributed to the procedure which is impossible following the procedure. It is unreasonable to expect that all the future inspection conditions could be predetermined at the time of the writing of the procedure (at least for many inspections), and thus the predetermined nature of the procedure conflicts with the varying inspection targets. (In the holistic framework this is explicitly included in "application parameters", but the conflict with the procedural approach is not stated.)

In summary, the recent advances in the holistic approach of inspection reliability have brought into light several long-standing handicaps of the current paradigm for inspection qualification. In essence, the paradigm states, that inspection reliability is the property of the inspection system or procedure, it may be measured in qualification or other ways and it may be reduced or compromised in actual inspections by unforeseen conditions (application parameters) or human factors. As stated above, this paradigm leaves both learning and adaptation outside the scope of the discussion (in practice, to be addressed when writing the next procedure or designing the next system). Excluding learning and adaptation from using a set procedure induces a combination of conflicting requirements to the inspectors and necessitates various "corrective" elements to move from the "system performance" to the "actual performance" as shown in the holistic approach.

1. Learning-centered paradigm for inspection performance

To address the limitations of the holistic approach, an alternate viewpoint on NDE reliability is developed. This is not meant to replace the holistic approach, but to complement it and to offer an alternate approach for improving NDE reliability. As a starting point, we take the viewpoint, that the expected NDE performance depends on the inspector (in the context of the available tools, i.e. inspection system, procedure and inspection target). Different inspectors may show different performance using the same procedure and inspection target, as is well known. Since procedure is the same, these differences result from differences in the inspector proficiency. To obtain optimal performance by different inspectors, the inspector performance and/or the procedure needs to be adapted to the situation at hand. Thus, equivalently, the inspector performance can be seen as result of inspector learning.

In summary, the expected inspector performance is a function of the available tools (procedures) and inspector adaptation (both to the tools and the task at hand). Successful adaptation is the result of inspector learning. Thus, primary way to improve inspection performance is to improve inspector learning to adapt the used procedure to existing conditions. Likewise, the primary way to monitor and control the expected performance is to test adaptation by testing performance.

Learning, in this context, is not necessarily beneficial. It is possible for inspectors to learn practices that lower the performance. Also, it does not necessitate conscious deliberation. It is expected that learning is continuous and on-going during the working life of the inspector. Thus, the expected performance will vary continuously and may improve or worsen. In fact, re-qualifications often fail which is notorious evidence of changing performance after qualification.

2. The necessary conditions for learning

NDT performance is now seen primarily as a result of inspector learning (i.e. acquired ability to adapt tools and behaviour to the task at hand for optimal performance). Improving performance is thus primarily a matter of improving the conditions for inspector learning. Likewise, estimating expected performance is primarily a matter of testing obtained ability in different settings. Consequently, it is of interest to study the necessary preconditions of learning and how they are present in current NDT inspection qualification framework. The necessary conditions for learning can be summarized as in Figure 2 [11,12].

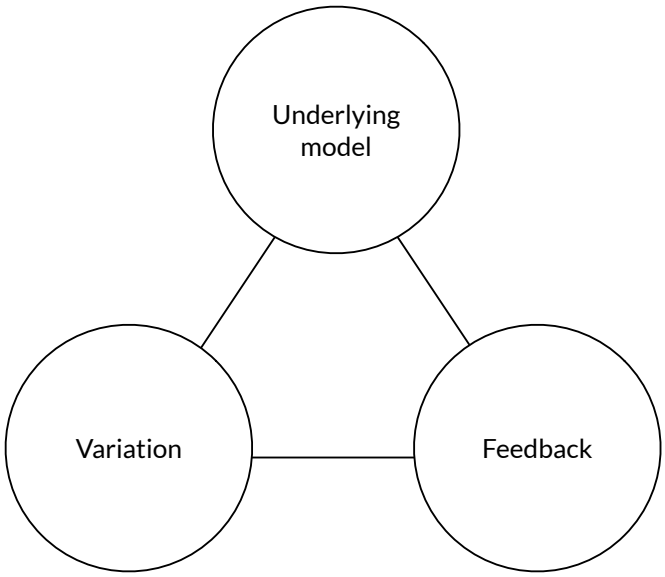

Fig. 2. Necessary conditions for learning adapted to the NDE context.

By comparing the necessary condition of learning with the holistic model of NDE reliability we note, that the models overlap significantly. Both include explicitly the significance of the underlying physical phenomenon (Intrinsic capability and Underlying model). The "Application parameters" correspond with the "Variation" in the learning model. Whereas the application parameters in the holistic model are seen to threaten the NDE reliability, variation is both a challenge and prerequisite for learning the necessary adaptation. The biggest difference is with the "human factor" and "organizational context" parts, which are implicit in the learning model. Feedback, a necessary part in the learning model is not represented in the holistic model. Thus, the role of feedback may be under-represented in the current viewpoint. Available feedback for inspectors is scarce within the current setting. The qualification exercise gives some limited feedback on pass/fail basis. Training also gives feedback on the signal-response of the used method from training samples. However, feedback on reliability is especially scarce.

3. Experimental study on the effect of feedback and variation on inspection reliability

As noted, the most significant difference to prevailing model is the significance of feedback and variation to inspection reliability. To test this hypothesis, a small empirical study was completed. A typical nuclear industry mock-up component (stainless steel tube butt-weld) was scanned with a phased-array ultrasonic inspection set-up. The mock-up contained three artificially induced thermal fatigue cracks of different size. The acquired data was analysed and the flaw signal was extracted using the eFlaw technology [13]. Flaw indications were then removed from the data to provide a clean defect-free data, where the extracted flaws could be re-introduced at various locations at will. An online tool was created to provide facilities for the inspectors to identify possible flaws. The online tool includes a simplified UT set-up with B-scan image of the data and software gain control and tools to indicate cracks with point and click (Figure 3.). The system generates random flawed data images on the fly by re-introducing extracted flaw signals at different locations in the data (also, the data is rotated and flipped to disable direct comparison of background noise and using the difference as hint for flaw detection). The user than analyses the image (possibly changes the gain to get more confidence on the results) and indicates found flaws by clicking them. After the image is analysed, the user requests next image by clicking a button. After 150 images have been analysed (many of them without flaws), the system uses the provided hits and misses to

compute a POD curve and confidence bounds. Since the mock-up contained only three cracks, the data provided limited variation of real cracks to the inspectors. To augment this, flaw amplitudes were artificially altered to produce a wider set of equivalent flaw sizes.

For this study, additional "learning" version of the tool was created. In this "learning mode", after user requests next image, the system shows results of the current images (i.e. hits, misses and false calls in the current image). This set-up thus provides the inspector with direct feedback of his success and should thus better facilitate learning this particular inspection task.

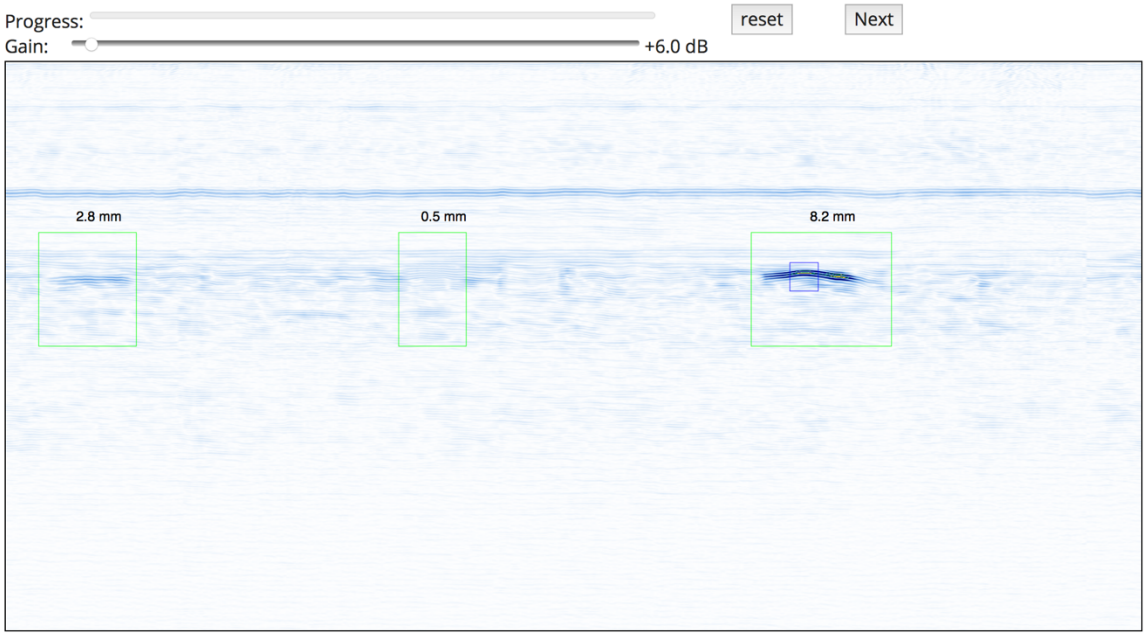

Fig. 3. The simplified inspection view used to gather hit/miss data in feedback mode. The green rectangles show true crack locations and equivalent sizes. Blue rectangle shows user-indicated crack location.

A simple amplitude threshold just above the highest noise peak in the data was used to generate a POD curve. This curve represents the performance attainable via a simple amplitude rule, optimized for this data and provides a useful reference to compare inspector performance. The POD curve is shown in Figure 4. The curve shows $a_{90/95}$ value of 3.75 mm.

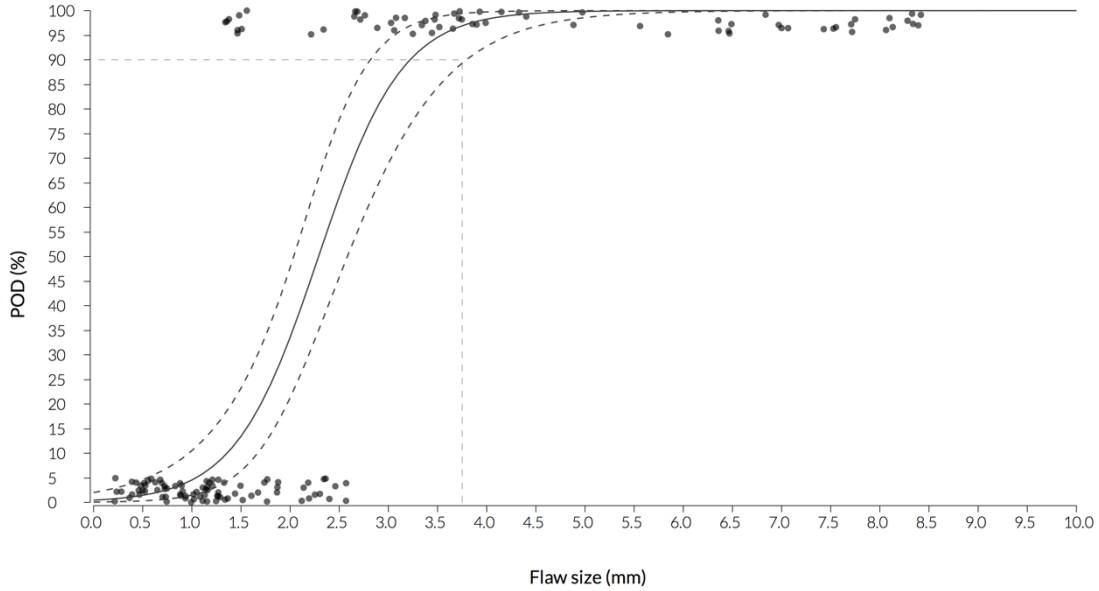

Fig. 4. POD curve generated from the automated system with a fixed amplitude threshold set just above the highest noise-peak in the data.

The set-up was tested during a level-III inspector training course. 9 students available were randomly divided in two groups (A and B). In the beginning of the course, all inspectors used the tool to get a base-line result before the course. After this baseline was recorded, group A continued training with the system in the normal mode during the week-long course. Group B had similar amount of training using the "learning mode". Finally, at the end of the course, final POD curve was obtained from all the inspectors. The final number of full training sessions varied as the students took different times to do the test. The students were not penalized for false calls. However, one person made so many false calls (1290), that they effectively obscured any information about true performance and thus this person was excluded from further study. The final numbers of the results are summarized in Table 1.

Table 1. Final number of full POD exercises

Inspector	Version	Number of full results
A	Non-learning mode only	1
B	Non-learning mode only	1
C	Non-learning mode only	1
D	Non-learning mode only	2
E	Learning-mode	3
F	Learning-mode	3
G	Learning-mode	3
H	Learning-mode	4

4. Results

The first-trial results are shown in Figure 5. Most of the inspectors reached better than reference (Figure 4) performance. False call rates varied significantly, and there was no clear correlation between false call rates and $a_{90/95}$. In Figure 6, the final results are presented, after learning trials and full POD trials as listed in Table 1. The learning-mode students show

significant improvement with one significant outlier. A more detailed analysis of the POD results for this inspector revealed, that the learning significantly improved the small hits. However, the number of "big misses" and the largest missed cracks were not decreased to the same degree, and thus the improvement did not translate into improved $a_{90/95}$ results.

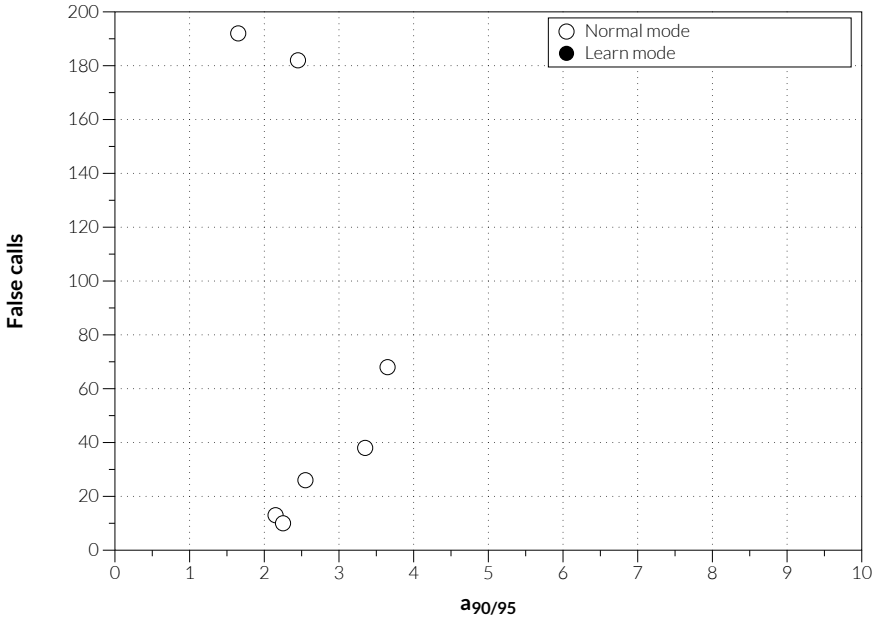

Fig. 5. First trial POD results from the students.

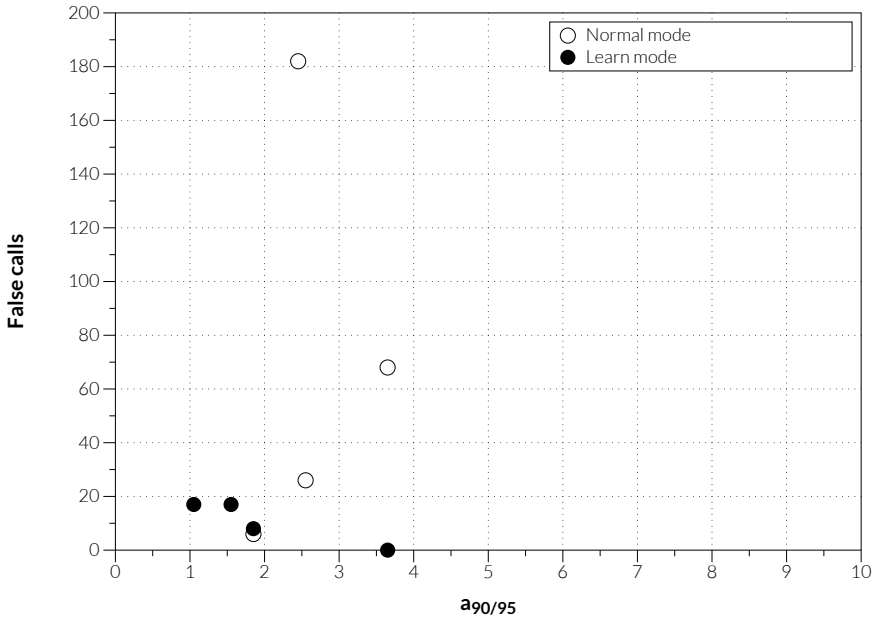

Fig. 6. Final trial POD results from the students. The learning-mode students show significant improvement with one significant outlier.

4. Conclusions

The limited evidence gathered during this study suggests, that direct feedback helps inspectors improve the reliability of flaw detection, as measured by the POD curve and $a_{90/95}$ values. This is to be compared with previous evidence showing no improvement on reliability

during normal work experience [9]. The results also indicate, that virtual flaws can be used successfully to give more direct feedback than previously available.

References

[1] Anon. 2012. Non-destructive Testing. Qualification and Certification of NDT Personnel. General Principles. ISO 9712:2012. International Organization for Standardization.

[2] Anon. 2016. Non-destructive Testing of Steel Forgings - Part 4: Ultrasonic Testing of Austenitic and Austenitic-ferritic Stainless Steel Forgings. EN 10228-4. European Committee for Standardization.

[3] Anon. 2007. The European Methodology for Qualification of Non-destructive Testing, Third Issue. ENIQ Report Nr. 31, EUR 22906 EN, ISSN 1018-5593.

[4] Lemaintre, P., Koblé, T. D. & Doctor, S. R. 1996. "Summary of the PISC Round Robin Test Results on Wrought and Cast Austenitic Steel Weldments, Part III: Cast-to-cast Capability Study". International Journal of Pressure Vessels and Piping (69) pp. 33-44.

[5] Anon. 2008. Nondestructive Evaluation Requirements for Fracture-Critical Metallic Components. NASA-STD-5009. National Aeronautics and Space Administration Washington, DC 20546-000. 28 pp.

[6] Anon. 2009. Nondestructive Evaluation System Reliability Assessment. Department of Defence Handbook. MIL-HDBK-1823A. 171 p.

[7] Cowfer, C.D. 1991. Basis / Background for ASME Code Section XI Proposed Appendix VIII: Ultrasonic Examination Performance Demonstration. Nuclear Engineering and Design, 131, pp. 313 - 317.

[8] Anderson, M., Diaz, A. & Doctor, S. 2012. Evaluation of Manual Ultrasonic Examinations Applied to Detect Flaws in Primary System Dissimilar Metal Welds at North Anna Power Station. PNNL-21546, Pacific Northwest National Laboratory, Richland, Washington. ADAMS Accession No. ML12200A216.

[9] Bertovic, Marija. 2016. Human Factors in Non-Destructive Testing (NDT): Risks and Challenges of Mechanised NDT. Bundesanstalt für Materialforschungn und -prüfung (BAM), Berlin, Germany. ISBN 978-3-9817502-7-0.

[10] Müller, C., Bertovic, M., Pavlovic, M., Kantzler, D., Ewert, U, Pitkänen, J & Ronneteg, U. 2013. Paradigm Shift in the Holistic Evaluation of the Reliability for NDE Systems. Materials Testing, 55 (4), pp. 261-269.

[11] Marton, F. 2015. Necessary Conditions for Learning. Taylor & Francis, New Your, U.S.A. ISBN 978-1-315-81687-6.

[12] Sterman, J. 2000. Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill Higher Education. ISBN: 9780072389159

[13] Koskinen, T., Virkkunen, I., Papula, S., Sarikka, T. & Haapalainen, J. 2017. Producing a POD Curve with Emulated Signal Response Data. Insight. Accepted for publication.