JRC Scientific and Technical Reports

ENIQ TECHNICAL REPORT

LINK BETWEEN RISK-INFORMED IN-SERVICE INSPECTION AND INSPECTION QUALIFICATION

ENIQ report No 36

Authors: B. Shepherd, L. Gandossi, K. Simola

ENIQ

European Network for Inspection and Qualification

JRC
EUROPEAN COMMISSION

Institute for Energy

EUR 23928 EN - 2009

Mission of the Institute for Energy
The Institute for Energy provides scientific and technical support for the conception, development, implementation and monitoring of community policies related to energy. Special emphasis is given to the security of energy supply and to sustainable and safe energy production.

European Commission
Directorate-General Joint Research Centre (DG JRC)
http://www.jrc.ec.europa.eu/

Institute for Energy, Petten (the Netherlands)
http://ie.jrc.ec.europa.eu/

Contact details:
Luca Gandossi
Tel: +31 (0) 224 565250
E-mail: luca.gandossi@jrc.nl

Legal Notice

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

The use of trademarks in this publication does not constitute an endorsement by the European Commission.

The views expressed in this publication are the sole responsibility of the author(s) and do not necessarily reflect the views of the European Commission.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server http://europa.eu/

EUR 23928 EN
ISSN 1018-5593
Luxembourg: Office for Official Publications of the European Communities

Printed in the Netherlands

European Commission
Directorate General Joint Research Centre
Institute for Energy
Petten, the Netherlands

ENIQ TECHNICAL REPORT:

LINK BETWEEN RISK-INFORMED IN-SERVICE INSPECTION AND INSPECTION QUALIFICATION

June 2009 ***ENIQ Report No 36*** ***EUR 23928 EN***

Approved by the ENIQ Steering Committee for publication

Documents published by ENIQ, the European Network for Inspection and Qualification, belong to one of the following 3 types:

Type 1 — **Consensus Document**
A *consensus document* contains harmonised principles, methodologies, approaches and procedures, and stresses the degree of harmonisation on the subject among ENIQ members.

Type 2 — **Position/Discussion Document**
A *position/discussion document* may contain compilations of ideas, expressions of opinion, reviews of practices, or conclusions and recommendations from technical projects.

Type 3 — **Technical Report**
A *technical report* is a document containing results of investigations, compilations of data, reviews and procedures without expressing any specific opinion or valuation on behalf of ENIQ.

This document "ENIQ Technical Report: Link between Risk-Informed In-Service Inspection and Inspection Qualification" (ENIQ Report No 36) is a Type 3 document.

FOREWORD

This report describes the work performed and presents the results from the group sponsored project “Link Between Risk-Informed In-Service Inspection and Inspection Qualification”. The project was coordinated by Doosan Babcock with a group of nuclear utilities providing financial sponsorship. A significant proportion of the work was performed by the EC Joint Research Centre (JRC) Petten under JRC funding. The main subcontractor was VTT.

The ideas developed in this work were originally discussed within the ENIQ Steering Committee. ENIQ, the European Network for Inspection and Qualification, was set up in 1992 in recognition of the importance of the issue of qualification of NDE inspection procedures used in in-service inspection programmes for nuclear power plants. Driven by European nuclear utilities and managed by the European Commission Joint Research Centre (JRC) in Petten, the Netherlands, ENIQ is intended to be a network in which available resources and expertise are managed at European level. ENIQ work is carried out by two sub-groups: the Task Group on Qualification (TGQ) focuses on the qualification of in-service inspection (ISI) systems, and the Task Group on Risk (TGR) focuses on risk-informed in-service inspection (RI-ISI) issues. More information on the ENIQ network and its activities can be found at http://safelife.jrc.ec.europa.eu/eniq/.

This report is essentially the same as that previously issued as a Doosan Babcock internal report (BWO Shepherd, L Gandossi, KA Simola, Link Between Risk-Informed In-Service Inspection and Inspection Qualification, Report No. TR-08-071, Project 79622). ENIQ has decided to publish this report as an ENIQ type-3 document, so that its content receives a more widespread circulation. Publication was approved by the ENIQ Steering Committee. The authors are BWO Shepherd (Doosan Babcock), L Gandossi (JRC) and KA Simola (VTT).

The voting members of the ENIQ Steering Committee are:

R Chapman	British Energy, United Kingdom
P Dombret	Tractebel, Belgium
K Hukkanen	Teollisuuden Voima OY, Finland
R Schwammberger	Kernkraftwerk Leibastadt, Switzerland
B Neundorf	Vattenfall Europe Nuclear Energy, Germany
J Neupauer	Slovenské Elektrárne, Slovakia
S Pérez	Iberdrola, Spain
U Sandberg	Forsmark NPP, Sweden
P Kopcil	Dukovany NPP, Czech Republic
D Szabó	Paks NPP, Hungary

The European Commission representatives in ENIQ are L Gandossi and T Seldis.

TABLE OF CONTENTS

SUMMARY

There is a growing need for a quantitative measure of inspection effectiveness as an input to quantitative risk-informed in-service inspection (RI-ISI). A Probability of Detection (POD) curve could provide a suitable metric. However there can be significant problems associated with generating realistic POD curves by practical trials. The ENIQ inspection qualification methodology can provide high assurance that an inspection system will achieve its objectives, but is not designed to provide a quantitative measure of the type that can be used in RI-ISI analysis.

A project was therefore set up with the following objectives:

- Investigate approaches to quantifying the confidence associated with inspection qualification.
- Produce guidelines on how to relate inspection qualification results, risk reduction and inspection interval.
- Apply the guidelines in practice via a pilot study, and modify them as required based on the experiences from the pilot study.
- Provide a forum for focused discussion and agreement on the approaches and guidelines.

The report discusses how a simplified POD curve, such as a step curve, could be used as the target for inspection qualification, or as an output from it. Work to investigate the sensitivity of relative risk reduction to the details of the POD curve is described from which it is concluded that use of a simplified POD curve could be justified.

Two methods for quantifying the outcome from inspection qualification are described. The first method is a relatively simple process based on direct expert judgement. Previous work to investigate the relationship between POD and margin of detection, which could facilitate this judgement, is discussed.

The second method is based on a more rigorous structured process employing Bayesian statistics, in which the subjective degree of belief in inspection capability derived from a Technical Justification (TJ) is expressed in probabilistic terms, and combined with data from practical trials results. Guidelines for the application of this Bayesian Methodology are provided in an Annex.

Two pilot studies are described which involved a qualification body applying the quantification methods in practice.

Work to model the relation between inspection qualification outcome, risk reduction and inspection interval is described, and an example of this process based on data from the 2nd pilot study is presented.

The main objectives of the project were achieved. Recommendations for further work to make the approaches developed more robust are provided.

1 INTRODUCTION

This final report describes the work performed and presents the results from the group sponsored project "Link Between Risk-Informed In-Service Inspection and Inspection Qualification". The project was coordinated by Doosan Babcock with a group of nuclear utilities providing financial sponsorship. A significant proportion of the work was performed by the EC Joint Research Centre (JRC) Petten under JRC funding. The main subcontractor was VTT.

The project officially started in June 2006 and had an initial planned duration of 15 months. An increase in sponsorship above the original target allowed the workscope to be extended with a final completion date of May 2008.

The sponsors were:

1) Forsmarks Kraftgrupp AB (representing a group of Swedish utilities)
2) TVO
3) VGB PowerTech Service GmbH
4) Iberdrola SA

The sponsors were kept informed of progress through periodic progress reports and meetings, and they had the opportunity to steer the direction of the project.

2 BACKGROUND

In-service inspection plays a key role in ensuring the continued safe and economic operation of nuclear plant. It is therefore very important that the effectiveness of the inspection system (combination of inspection equipment, procedure and personnel) is properly understood and achieves the intended performance in practice. In Europe, this inspection qualification is widely performed in accordance with the European Network for Inspection and Qualification (ENIQ) methodology [Ref. 1].

The inspection locations, frequencies and methods have traditionally been based primarily on the type and safety category of the equipment determined from the original design basis. However improvements in probabilistic safety assessments combined with many years' operating experience have resulted in increasing interest in the adoption of risk-informed in-service inspection (RI-ISI). ENIQ has recently issued a discussion document which provides a general framework for RI-ISI [Ref. 2].

Although consequence of failure is not influenced by inspection, the probability of failure is. The effectiveness of inspection is therefore an important input for RI-ISI analysis. If a quantitative RI-ISI analysis is to be performed, then a quantitative measure of inspection effectiveness is needed in order to calculate the reduction in risk associated with inspection. A Probability of Detection (POD) curve would provide ideal data.

POD data is normally generated by performing practical trials on a large number of defects in test pieces. The probability of detection is then plotted against an appropriate defect parameter. From statistics, a minimum of 29 defects with the same parameter all need to be detected in order to establish a 90% probability of detection at a lower bound confidence of 95%. This provides just one point on the POD curve.

It can be relatively straightforward to establish POD curves for NDT methods such as magnetic particle or dye penetrant testing, where the main parameter influencing detection is generally defect length. However in the case of ultrasonic inspection (which is the main NDT method for ISI of nuclear plant) there are many defect variables which affect detectability, such as defect depth, length, location, tilt, skew, shape, and roughness.

In recognition of these limitations, the ENIQ approach to inspection qualification is based on a combination of technical justification and test piece trials. Technical justification involves assembling supporting evidence for inspection capability (results of capability evaluation exercises, feedback from site experience, theoretical models, physical reasoning). The balance between the various elements and the level of detail involved in qualification are judged separately for each case.

The output from the ENIQ qualification process is generally a statement concluding whether or not there is high confidence that the required inspection capability will be achieved in practice, for the specified inspection system, component and defect range. However the ENIQ methodology is not designed to provide a quantitative measure of inspection capability of the type which can be used by quantitative RI-ISI. It also means it is difficult to "benchmark" the confidence associated with any given inspection qualification.

This group sponsored project was therefore set up to investigate and demonstrate approaches which provide some objective measure of the confidence which comes from inspection qualification, and which allow risk reduction associated with a qualified inspection to be calculated.

3 OBJECTIVES

The objectives of the project as defined in the project proposal [Ref. 3] were to:

- Investigate approaches to quantifying the confidence associated wth inspection qualification.

- Produce guidelines on how to relate inspection qualification results, risk reduction and inspection interval.

- Apply the guidelines in practice via a pilot study, and modify them as required based on the experiences from the pilot study.

- Provide a forum for focused discussion and agreement on the approaches and guidelines.

4 OVERVIEW OF WORKSCOPE

The main elements of the work were:

- Investigate the concept of using a simplified POD curve to represent inspection performance, including a study of the sensitivity of risk to level of detail in the POD curve

- Develop draft guidelines on how to use direct expert judgement quantify inspection performance as a POD during qualification

- Develop draft guidelines on how to use a Bayesian statistics approach to quantify inspection performance as a POD during qualification

- Perform pilot studies using these draft guidelines to quantify the outcome from qualification of two technical justifications. Update the guidelines to take account of lessons learnt from the pilot studies.

- Illustrate how, if defect growth can be modelled, it is possible to link inspection qualification results, risk reduction and inspection interval.

- Issue final report and guidelines and provide recommendations for any further work.

Each of these is described in turn in the following sections.

5 ADOPTION OF SIMPLIFIED POD CURVE

As explained above, it is generally not practicable to produce a POD curve for specific combinations of plant component and UT inspection system based on experimental data, and the ENIQ process is not designed to produce such a curve. However it is possible to define a POD curve which, if it were met in practice, would achieve a given level of risk reduction. It might then be possible to use this user-defined POD curve as the target for qualification. The task of the qualification body would then be to judge whether or not this POD curve could be considered a lower bound for actual inspection capability.

Alternatively, no target POD might be specified beforehand, and the inspection qualification body is asked to quantify inspection performance in terms of POD.

It will be easier for the qualification body to make this judgement the simpler the shape of the POD curve, regardless of whether it is provided as a target, or whether it is required as an output. The simplest case will be a step function where a fixed POD is assumed for defects above the qualification size, and no detection capability assumed for defects below that size. It will also be easier the lower the POD required.

However the risk reduction associated with an inspection will tend to increase if credit is taken for detecting defects below this “cut-off” qualification size, and will also tend to increase the higher the POD for a given defect size. It is therefore useful to understand the sensitivity of risk reduction to the level of detail in the POD curve.

This was studied by JRC and the work is reported in detail in [Ref. 4]. A summary of the work follows.

In broad engineering terms, the risk of failure of a component (e.g. a weld) is given by

$$\text{Risk}=pof \times cof \qquad (1)$$

where pof is the probability of failure and cof the consequence of that failure, expressed in some sort of convenient metric.

It is assumed that the inspection system under consideration has been designed to target a specific acting degradation mechanism leading to crack-like defects. The inspection programme will only affect the probability of failure. As an inspection is carried out, some knowledge regarding the (previously uncertain) state of the plant is gathered, and the probability of failure is (usually) reduced. If defective components are found, these are assumed to be repaired or removed. The consequence of failure, depending on many factors such as plant layout, presence of redundant or mitigating systems, etc. will not be affected by the inspection. It is therefore straightforward to see how the risk will change before and after an inspection is carried out by simply considering the change in probability of failure.

A risk reduction percentage, R, can be defined as:

$$R=(1-\frac{pof_{with}}{pof_{without}})\cdot 100 \qquad (2)$$

where $pof_{without}$ is the probability of failure without inspection, and pof_{with} is the probability of failure with inspection. Values of *R* close to 100 mean a significant risk reduction due to the inspection. A perfect inspection (one capable of finding all defects of all sizes) will reduce the probability of failure pof_{with} to zero, and therefore *R* will be equal to 100. On the other hand, values of *R* close to zero mean a small change in risk, due to a poor inspection.

In a detailed, fully quantitative approach to RI-ISI, the inspection programme would be modelled within a probabilistic structural integrity model. In this case, extensive knowledge or assumptions are needed about materials, loadings, initial crack distribution, crack growth behaviour, probability of detection, etc. Typically, a Monte Carlo analysis of the problem would be performed, repeating many times over a deterministic structural integrity assessment, each time sampling the required input parameters from the appropriate distributions and verifying whether the structure has failed or is in a safe state for that particular combination of input variables. The main drawback is the much higher requirement in terms of resources such as time, calculating power, need to know (or to make assumptions about) material properties, loadings, etc.

A simplified approach was therefore considered and the possibility of crack growth was ignored.

An inner surface breaking crack in the component is postulated. The through-wall extent of such crack, a, (also called “depth”) is assumed to be a random variable and is normalised with respect to the component thickness so that it takes any value between 0 and 1.

The following functions are then considered:

- A probability distribution of flaw size, $\lambda(a)$
- A function expressing the probability of failure as a function of flaw size, $\phi(a)$
- A POD curve, *p(a)*

If the functions $\lambda(a)$, $\phi(a)$ and *p(a)* are defined, it is straightforward to calculate the probability of failure without inspection, $pof_{without}$, by integrating $\lambda(a) \times \phi(a)$ over the flaw size:

$$pof_{without} = \int_0^1 \lambda(a)\phi(a)da \tag{3}$$

And, since 1-*p(a)* is the probability of missing a defect of size *a*, the probability of failure with inspection, pof_{with}, is given by:

$$pof_{with} = \int_0^1 \lambda(a)\phi(a)(1-p(a))da \tag{4}$$

The risk reduction percentage expressed in Equation 2 can therefore be evaluated and was used to study the influence of different types of POD curve on risk reduction, as a function of defect distribution and material toughness (sensitivity of failure probability to flaw size). Note that in this study the number (or frequency) of cracks has no effect, since it would increase pof_{with} by the same percentage as $pof_{without}$, and *R* would remain unchanged.

It should also be noted that in this approach it is assumed that the probability of failure, $\phi(a)$, and the probability of detection, *p(a)*, are functions of crack depth only.

Although there is a lack of information on the distribution of defects in nuclear plant, the limited information available suggests an exponential distribution. Figure 1 below illustrates some defect distributions modelled by the truncated exponential function. Since Ref. 4 describes this function, it is sufficient here to note that the different distributions have been derived by varying its parameter μ. A small value for μ represents defects being mainly at the small end of the spectrum whereas as μ increases the distribution becomes more uniform.

Probability of failure was also modelled, in this case using the cumulative Beta distribution and varying parameters α and β (see Ref. 4). Figure 2 below illustrates three distributions corresponding to different parameters. Selecting α and β both equal to 1 represents a case where probability of failure increases in direct proportion to crack depth. The two other cases illustrated represent a situation where probability of failure becomes high even for relatively small cracks (brittle material) and where probability of failure remains low until crack size is relatively high (tough material).

The third function modelled was the POD curve. For this study, simple functions were used to approximate the shape of the POD curve. A first convenient approximation is to suppose that the POD is zero for all crack sizes *a* below a certain threshold depth a_{th}, and equal to a plateau value p_{pl} for $a > a_{th}$. In Figure 3, two such curves are illustrated, the first defined for a_{th}=0.2 and p_{pl}=0.9, and the second for a_{th}=0.4 and p_{pl}=0.8.

A second type of POD curve which can be considered is only slightly more sophisticated, showing a sloped transition between p=0 and p=p_{pl} (and thus requiring for complete definition a third parameter, a_{pl}, i.e. the crack depth at which p=p_{pl} is attained). Figure 3 also includes one example of such a curve (a_{th}=0.6, a_{pl}=0.7 and p_{pl}=0.7). This latter type of curve was used to investigate the value added by having a slightly more informative POD curve.

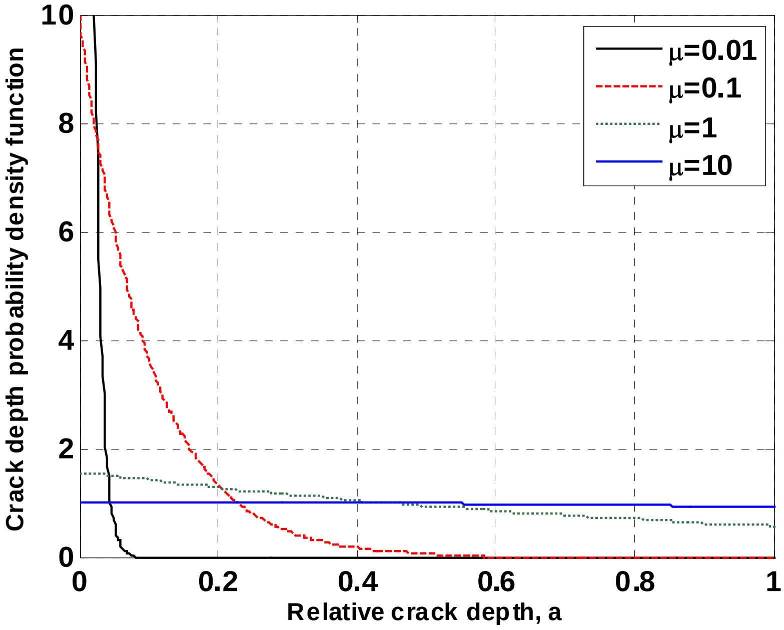


Figure 1 Defect size distributions

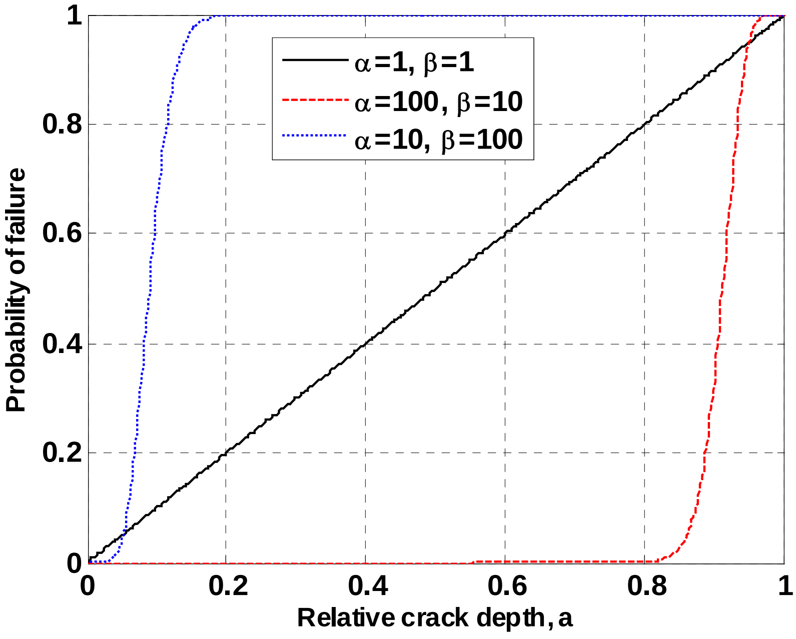


Figure 2 Probability of failure distributions

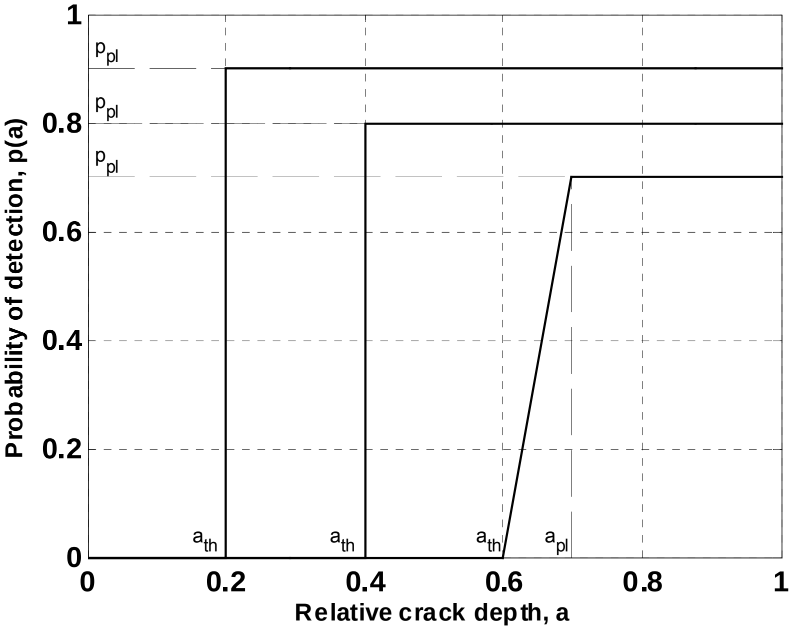


Figure 3 Simplified POD curves

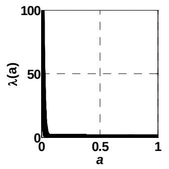
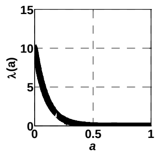
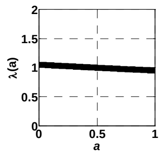
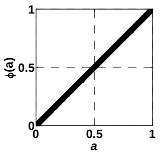
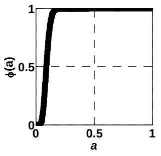
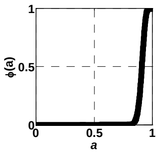
		Flaw depth distribution, $\lambda(a)$		
		μ=0.01	μ=0.1	μ=10
Probability of failure, $\phi(a)$				
α=1, β=1		CASE 1a	CASE 1b	CASE 1c
α=10, β=100		CASE 2a	CASE 2b	CASE 2c
α=100, β=10		CASE 3a	CASE 3b	CASE 3c

Figure 4 Nine base cases for POD sensitivity investigation

Having defined the functions representing defect distribution, probability of failure and POD, the risk reduction R can be determined using equations (2), (3) and (4).

Risk reduction was calculated first of all for the case of a step-curve POD ($a_{th} = a_{pl}$), for each of the 9 cases illustrated in Figure 4.

The risk reduction as a function of a_{th} (defect qualification size) and p_{pl} (maximum POD) was then calculated for each of these base cases (POD is a step curve).

Ref. 4 presents the full set of results but one example is provided in Figure 5 below. This presents the results from base case 3a which is possibly the most representative of material in nuclear plant.

Curves such as these can be used to make a choice on how the resources allocated to the design and qualification of an inspection system could be invested. Different strategies could be envisaged:

- For a given, fixed a_{th}, (which could be seen as the target defect qualification size) determine which level p_{pl} should be used to obtain the desired risk reduction;
- For a given, fixed p_{pl}, determine which a_{th} should be used to obtain the desired risk reduction;
- A combination of the two above, maximising the risk reduction for the given resources (it could be the case that it is much easier – and cheaper – to push down a_{th} rather than push up p_{pl}, and still obtain the same risk reduction.

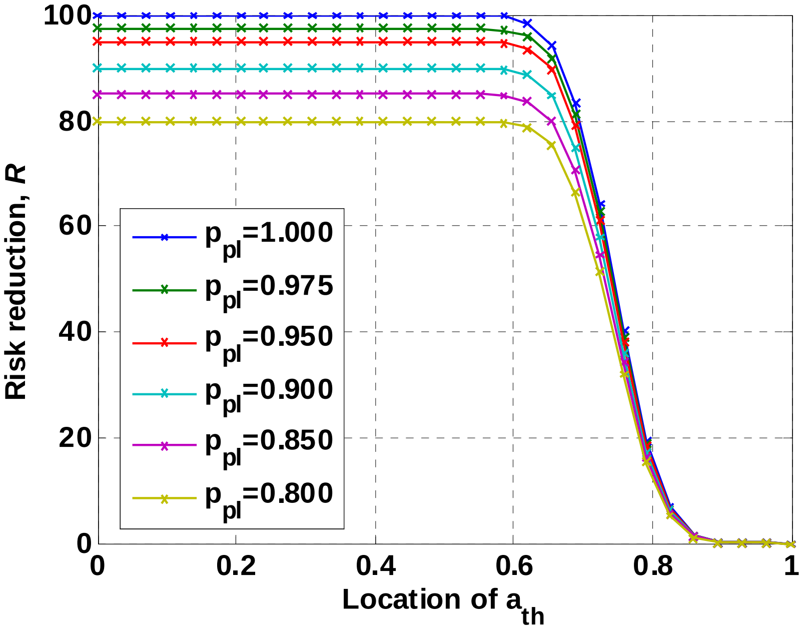


Figure 5 Sensitivity of R to POD parameters, base case 3a

It is important to note that the risk reduction R is a relative measurement, comparing risk with and without inspection. When R is close to 100 a significant risk reduction is achieved. When R approaches zero only a small change in risk is achieved. It does not on its own provide an insight into absolute risk reduction. For example reducing risk by 50% for case 2c (brittle material, significant proportion of large defects present) might be more valuable than reducing risk by 99% for case 3a (tough material, only small defects present).

The sensitivity of risk reduction to the level of detail in a POD curve was also investigated (the examples above were based on a simple step curve.) This was done by considering how risk reduction is affected by different slopes of the POD curve from its plateau value at a_p down to its zero value at a_{th} (see Figure 3 and the preceding text).

Figure 6 plots the results for base case 2b. Risk reduction is plotted for various values of a_{pl} (crack size at which plateau is reached) and various values of p_{pl} (plateau value of POD) against a_{th} (crack size at which POD becomes zero). For a given a_{pl}, the slope of the POD curve is more gradual the lower the value of a_{th}.

This figure illustrates how risk reduction "credit" can be taken for detecting defects below a_{pl} for this particular case whereas approximating the POD as an abrupt step curve (a_{th} = a_{pl}) results in a lower calculated risk reduction. As expected, this risk reduction is higher, the lower a_{pl} is and the higher p_{pl} is (detection of smaller defects, higher POD at plateau).

Figure 7 shows the results for base case 3a (expected to be most realistic for material in nuclear plant). In this case the sensitivity of risk reduction is quite different. Unless a_{pl} is quite high (large defect qualification size) there is no advantage in taking into account detection capabilty for defects below a_{pl} and in fact risk reduction does not even appear to depend on the value of a_{pl}. It only depends on p_{pl} (POD at plateau).

This approach to studying the sensitivity of risk reduction to the level and detail in the POD curve can be a very useful means of establishing an appropriate balance between risk reduction on the one hand, and avoiding placing excessive demands on the inspection system and expressing its capability as a detailed POD curve, on the other.

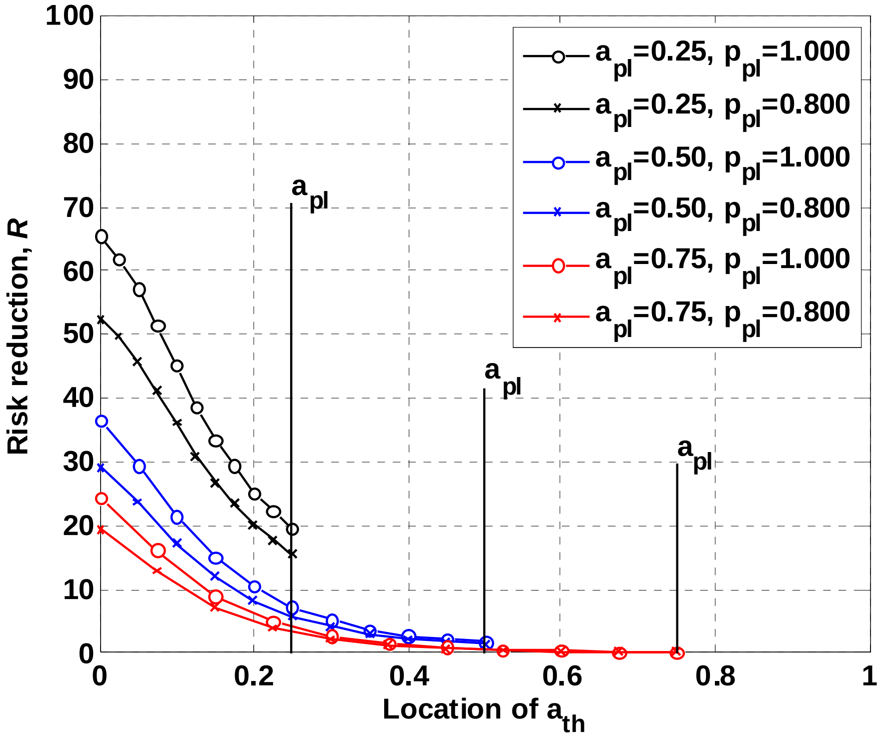


Figure 6 Sensitivity of risk to POD slope, base case 2b

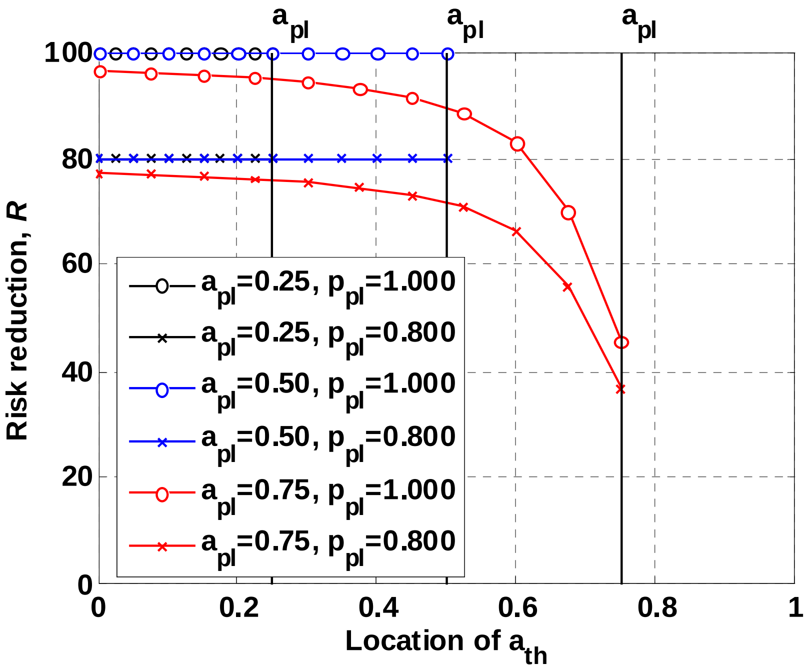


Figure 7 Sensitivity of risk to POD slope, base case 3a

6 DIRECT EXPERT JUDGEMENT OF POD

As discussed in the introduction to Section 5, it may be possible to define a simple POD curve as the target for inspection qualification ("user-defined" POD curve), or to request a simple POD curve as the output from qualification. The task of the qualification body (QB) will be easier, the simpler the shape of the curve and the lower the POD required. The approach described in Section 5 can be used to explore the sensitivity of risk to simplifying the POD.

The POD might be a simple step curve, with POD defined as some uniform level above the defect qualification depth, and taken to be zero below it. Figure 8 illustrates an example of this where the defect qualification size is 8mm and the target POD above this is 80%.

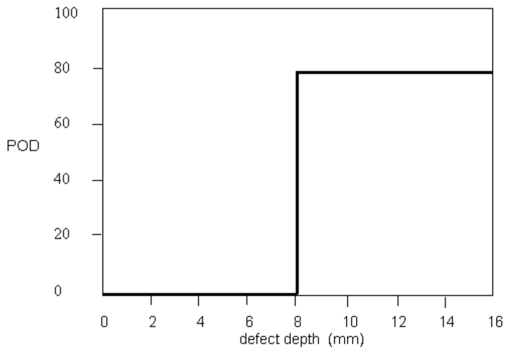


Figure 8 User Defined POD as Step Curve

In this case the qualification body would be asked to judge whether, on the basis of the qualification evidence provided, the inspection system should detect at least 80% of defects exceeding 8mm depth (and within the specified range of orientations, locations, shapes etc.)

The higher the POD level, the greater the risk reduction associated with the inspection. However as the target POD level increases, it may become increasingly difficult for the inspection system to meet that target. It is also likely to become increasingly difficult for the inspection qualification body to judge whether that target is achieved.

Even if the qualification body does not feel able to confirm that the inspection system will meet the target POD, it would still be useful to know what POD the QB agrees can be considered a lower bound. This could be determined by establishing first of all whether there was consensus among the QB members that the POD exceeded some low level such as 50% (for such an extreme case it would be hoped that this consensus would be reached immediately). The QB would then establish whether there was consensus that the POD exceeded some higher level, such as 60%, and so on until some POD level was reached which the QB felt unable to confirm was met. This process is illustrated in Table 1 below which also allows the confidence associated with the result to be recorded.

If the QB is 80% confident that at least 50% of the defects meeting the defect specification will be detected by the inspection system described, then a tick is placed opposite “50%”. The QB then moves to the second row. If it is 80% confident that at least 60% will be detected then a tick is also placed opposite “60%” and so on. When a level is reached where the QB considers it is not possible to make a judgement then a question mark is placed opposite it.

The “95% confidence” column is completed in the same way, but in this case based on a confidence level of 95%.

The Table could be completed by the QB members working as a team, or they could complete the Table independently and then compare and discuss their results. The discussion might influence some QB members to change their scores, after which they either attempt to produce a single set of results representing the QB as a whole (e.g. by averaging individual scores) or the QB reports the individual scores. It would then be the responsibility of the risk analyst to decide whether e.g. to average or use the most pessimistic value.

POD	80% confident that inspection exceeds (example)	95% confident that inspection exceeds (example)
50%	√	√
60%	√	√
70%	√	√
80%	√	√
90%	√	√
95%	√	?
98%	?	?

Table 1 Example of Direct Judgement Table

The higher the POD value used as the target for qualification, the more difficult it will be for the qualification body to be confident that the inspection system meets this target. Lowering the target will simplify matters, but at the expense of reducing the risk reduction which can be credited. The main problem facing the qualification body is in knowing how to relate evidence on inspection capability to a POD value. There might be evidence from trials, modelling and physical reasoning that all defects (within the prescribed range) should be above the reporting threshold and well above the noise level but does this mean the POD for such defects is at least 90%?

In order to help make this judgement, the work described in Ref. 5 has been performed to relate both signal to noise (S/N) ratio, and signal above report threshold, to POD (defect detection rate is a more correct term since the defect sample sizes were relatively small). The work involved a series of blind manual and automated UT trials on test pieces containing a range of artificial defects which provided a range of signal to noise, and signal to report threshold ratios. Relationships were established between these margins of detection, and “POD”. For manual UT, the relationship was that presented in Table 2.

Note that the artificial defects and test pieces did not need to be realistic simulations of actual defects and components since they were only required to act as sources of various levels of signals. A benefit of this approach is that unlike a POD curve the data generated from these trials may be generic, i.e. applicable to a broad range of inspection procedures, components and defect types.

S/N mean	Hits	Misses	"POD"
0 – 2.4	-	-	-
2.5 – 4.9	-	-	-
5.0 – 7.4	7	14	33
7.5 – 9.9	32	24	57
10.0 – 12.4	37	19	66
12.5 – 14.9	24	4	86
15.0 – 17.4	7	0	100
17.5 – 19.9	-	-	-

Table 2 Relationship between S/N and POD (manual UT)

As an example of the way in which these results could be used in practice, the proposed RI-ISI programme might be based on the assumption that defects above 15% wall thickness will be detected with a POD of at least 80%.

The role of the inspection qualification body would then be to make a judgement on the detectability of defects which exceeded this depth, and which were within certain other ranges determined by metallurgical considerations (covering tilt range, skew range, roughness, shape, location etc.)

Based on some combination of practical trials, theoretical modelling, physical reasoning and previous experience, the qualification body might conclude that all such defects should be detected with signal at least 14dB above noise.

Using the relationship suggested by Table 2, there would then be supporting evidence that the POD for such defects is around 86%, and the inspection system should therefore achieve the target POD.

Note that this simple example is for illustrative purposes only. Caution should be exercised when applying the results of Reference 5 since the defect population sizes were relatively small and POD can depend on more than just signal to noise ratio or signal above report threshold. For example in the case of automated UT it can also depend on pattern recognition. Further work to put the results of Reference 5 on a more conclusive footing is recommended.

7 BAYESIAN METHODOLOGY

In order to quantify the ENIQ qualification methodology, a Bayesian approach was proposed [Ref. 6] where the level of "confidence" in the technical justification is expressed in probabilistic terms. Bayesian statistics is widely applied in many scientific and technological fields. In this framework, it is particularly appealing because it offers a formal way of treating subjective probabilities (i.e. expert judgment).

According to the Bayesian interpretation, probability is a subjective degree of belief about events of interest based on the available evidence. In a Bayesian framework, the initial degree of belief regarding an unknown (or uncertain) variable is represented with a prior distribution. If there is no relevant evidence available, one should choose a so-called non-informative prior, which has minimal effect relative to data/experiment on the final inference. The prior distribution is then updated as new evidence is collected, using the Bayes' theorem. The resulting distribution is called the posterior distribution. When enough experimental evidence is available, it overwhelms the choice of prior and the results of Bayesian and frequentist analyses converge.

In the model, a population of defects is considered, characterised by a single, fixed flaw size, a. All the variables, such as component type and material, acting damage mechanism, the defect attributes (morphology, etc.) and the NDT system that is to be applied (procedure, equipment and personnel), are assumed to be defined (i.e. they need not to be specified). A value of the probability of detection, p, intrinsic to the NDT system under consideration (seen as the combination of all the above-mentioned variables) must exist. In the frequentistic interpretation of probability, p can be seen as the percentage of detection of that given class of defects, i.e. the number of detected defects divided by the total number of trials, as the total number of trials approaches infinity.

A trial in this context is considered as an opportunity to detect a defect, e.g. if there are 50 defects in the component or test piece inspected, or if there are 50 components or test pieces inspected and each contains one defect, the number of trials would be 50.

In the Bayesian statistical framework, it is assumed that the results of test piece trials are treated as a sample from a binomial distribution with parameter p. p is an unknown and fixed number in classical and frequentistic statistics, whereas it is considered as a random variable (with an associate probability distribution) in Bayesian statistics. It is natural to choose the Beta distribution to model the uncertainty related to p. The Beta distribution is conjugate to the binomial distribution and that renders the calculation of posterior distributions particularly simple. Thus:

$$p \sim Beta\ (\alpha, \beta) \tag{5}$$

The parameters α and β determine the shape of the Beta distribution. The distribution is defined for $\alpha>0$ and $\beta>0$. If $\alpha>1$ and $\beta>1$, the Beta distribution is unimodal. In this case, the expected value, $E(p)$, is given by $\alpha/(\alpha+\beta)$ and the mode by $(\alpha-1)/(\alpha+\beta-2)$. If $\alpha = 1$ and $\beta = 1$, the Beta distribution becomes the uniform distribution.

Figure 9 illustrates two examples of a Beta function to clarify what it represents. One has $\alpha = 1$ and $\beta = 1$ which is the situation where there is no prior knowledge regarding the value of *p.* If *p* represents POD, then in the absence of any other information, it is as likely that POD is 0% as it is to be 100%. The probability distribution is thus uniform. The other curve has $\alpha = 10$ and $\beta = 2$. In this case POD is likely to be around 95%, but could still be higher or lower as determined by its probability distribution.

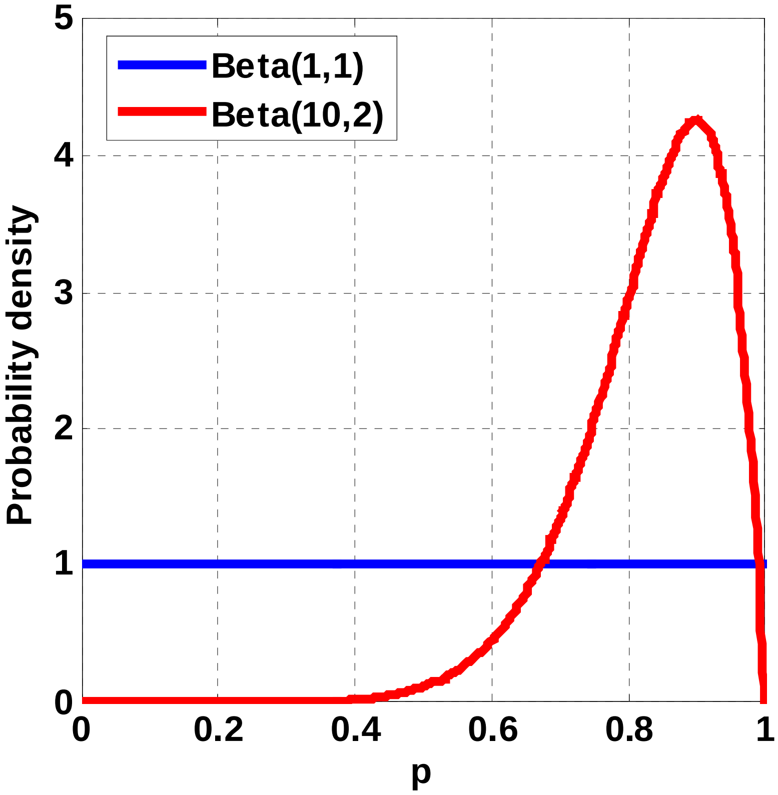


Figure 9 **Examples of Beta distributions**

In the Bayesian framework, the process starts by expressing the prior knowledge regarding p in the following way:

$$p \sim Beta\ (\alpha_{prior},\ \beta_{prior}) \tag{6}$$

If nothing is known about *p* before, a reasonable choice for the prior parameters would be α_{prior} = 1 and β_{prior} = *1*, expressing the fact that *p* could be anywhere in the interval [0, 1] with equal probability.

The second step of the process consists of gathering evidence regarding *p*. The most natural way to do so would be to carry out a number *N* of practical trials. In doing so, let us assume that the number of successes is N_s and the number of failures is N_f, so that N=N_s+N_f. The advantage of choosing a Beta distribution for *p* comes into play when determining the posterior. It can be shown that the posterior distribution is simply obtained as follows:

$$\begin{aligned} \alpha_{post} &= \alpha_{prior} + N_s \\ \beta_{post} &= \beta_{prior} + N_f \end{aligned} \tag{7}$$

At this stage, knowledge regarding *p* is fully expressed by

$$p \sim Beta(\alpha_{post}, \beta_{post}) \tag{8}$$

An attractive property of the updating process is that if new evidence becomes available, it can be readily used to obtain a second posterior. Notably, the order in which the first and second sets of trials are carried out does not affect the outcome.

Figure 10 illustrates how the Beta function (representing the probability distribution for POD) changes as is updated to take account of additional information.

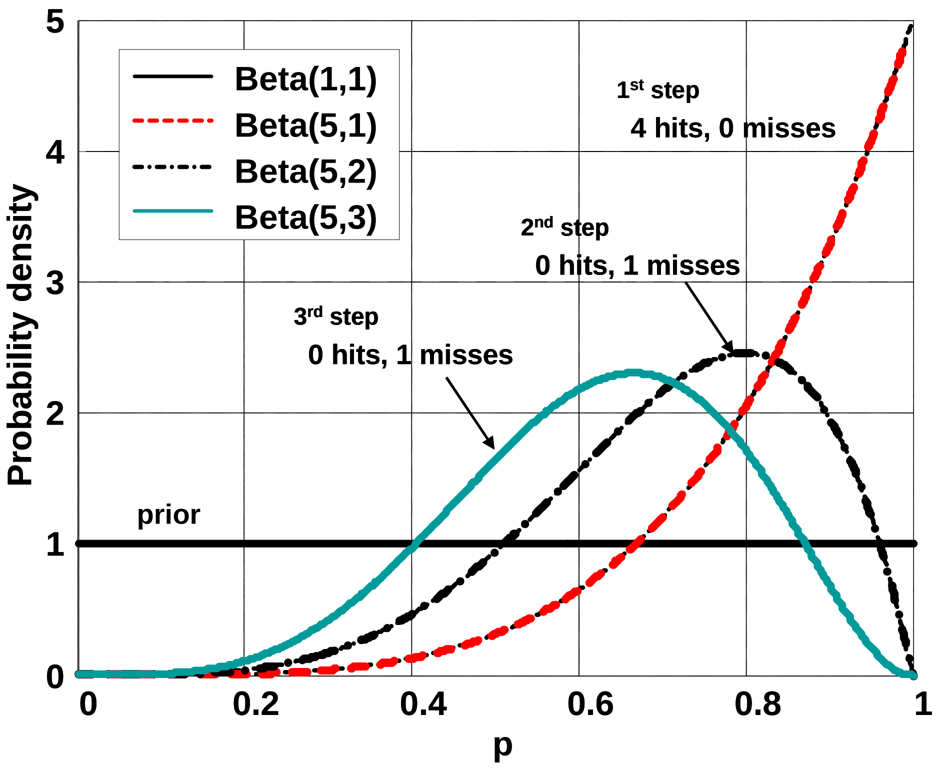


Figure 10 Examples of Beta function updating

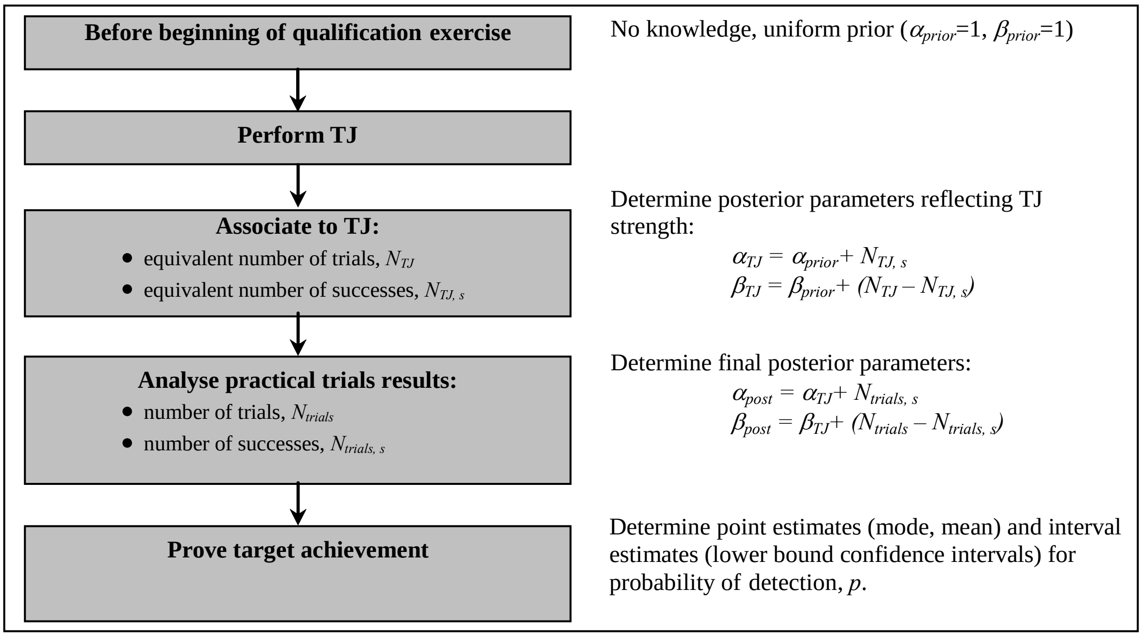
The essential idea proposed in [Ref. 6] follows naturally from these considerations and consists in interpreting the TJ in terms of an equivalent set of practical trials. It is suggested that the TJ be quantified using two numbers: an equivalent total number of trials, N_{TJ}, and an equivalent number of successes, $N_{TJ,\ s}$. For example it might be considered that a TJ provided the same degree of confidence in the inspection system as the detection during actual inspections of 19 out of 20 defects (which were within the defect qualification specification). In this case the TJ would be quantified as N_{TJ} = 20 and $N_{TJ,\ s}$ = 19.

These numbers, provided by qualification experts in a documented and transparent manner, could in principle be used in combination with the number of practical trials, N_{trials}, and associated number of successes, $N_{trials,\ s}$ to support the final statement that the NDE system can be considered qualified. Figure 11 schematically illustrates this idea.

The posterior distribution for *p*, expressed in Eq. (7), would thus be defined by the following parameters:

$$\begin{aligned} \alpha_{post} &= 1 + N_{TJ,s} + N_{trials,s} \\ \beta_{post} &= 1 + (N_{TJ} - N_{TJ,s}) + (N_{trials} - N_{trials,s}) \end{aligned} \tag{9}$$

The problem is now to assign meaningful values to the quantities N_{TJ} and $N_{TJ,\ s}$. In [Ref. 6] different approaches to tackle this issue were proposed. The discussion was further expanded (and complemented with several worked examples) in [Ref. 7].



Before beginning of qualification exercise

No knowledge, uniform prior (α_{prior}=1, β_{prior}=1)

Perform TJ

Associate to TJ:

- equivalent number of trials, N_{TJ}
- equivalent number of successes, $N_{TJ,\ s}$

Determine posterior parameters reflecting TJ strength:

$$\alpha_{TJ} = \alpha_{prior} + N_{TJ,\ s}$$
$$\beta_{TJ} = \beta_{prior} + (N_{TJ} - N_{TJ,\ s})$$

Analyse practical trials results:

- number of trials, N_{trials}
- number of successes, $N_{trials,\ s}$

Determine final posterior parameters:

$$\alpha_{post} = \alpha_{TJ} + N_{trials,\ s}$$
$$\beta_{post} = \beta_{TJ} + (N_{trials} - N_{trials,\ s})$$

Prove target achievement

Determine point estimates (mode, mean) and interval estimates (lower bound confidence intervals) for probability of detection, p.

Figure 11 Principles of the proposed Bayesian framework the quantitative modelling of the ENIQ methodology

The following basic principles were postulated to apply:

- If some evidence is missing, this should imply less weight for the TJ posterior. This means that the equivalent TJ sample size, N_{TJ}, should be smaller than in the case of stronger evidence.

- If evidence is present showing that some defects could be missed, this should imply a lower expected value of the TJ posterior, i.e. the ratio of $N_{TJ,s}$ over N_{TJ} should be smaller than in the case where the evidence is more convincing regarding detection capability.

It was also assumed that the TJ can be broken down into a number of elements and that the impact of each element towards demonstrating inspection capability is independent from the others. These elements could be for instance theoretical modelling experimental evidence, parametric studies, equipment considerations, data analysis, etc.

8 RELATIONSHIP BETWEEN POD, RISK REDUCTION AND INSPECTION INTERVAL

In section 5.1 the sensitivity of risk, or failure probability, to POD was discussed without taking into consideration the initiation and growth of flaws. In this section, the POD sensitivity studies are extended to take into account the inspection intevals. The studies are performed with an approach that integrates probabilistic fracture mechanics (PFM) calculations with a discrete time Markov process analysis to model piping degradation states at inspections, and to account for flaw and leak detection probabilities. Similar analyses could be performed with other probabilistic fracture mechanistic tools as well, as long as they have an option to account for inspection reliability and intervals.

8.1 Approach for the Probabilistic Fracture Mechanics Simulations

In the examples of this section, the probabilistic crack growth analyses were carried out with a modified version of fracture mechanistic analysis code VTTBESIT, developed by the Fraunhofer-Institut für Werkstoffmechanik (IWM), Germany and by VTT. With the VTTBESIT code it is possible to quickly compute the stress intensity factor values along the crack front and based on this simulation data the crack growth. VTTBESIT was modified by adding probabilistic capabilities to the code, which was originally intended for deterministic fracture mechanics based crack growth analyses.

In the analysis with the probabilistic version of VTTBESIT, the following randomised input parameters were used: exponential distribution for initial crack depth, exponential distribution for initial crack length and Poisson distribution for thermal load cycle frequency. The amount of crack growth in each time step is calculated from the respective crack growth (for instance fatigue or stress corrosion cracking) equation. The simulation ends either when the crack depth reaches the outer pipe surface, or the time cycles reach the end of plant lifetime. For each analysed case, hundreds or even thousands of separate simulations are calculated, and for each of these values of the above mentioned distributed input data parameters/variables are sampled at random from the respective probabilistic distributions. Each run is a 60-year simulation with the crack depth calculated at 1-year intervals, conditional on the existence of an initial flaw. The annual crack depth information for each simulation is tranferred to the second phase of the analyses.

In the second phase, the analyses are based on Markov processes. The Markov process is a stochastic process in which the probability distribution of the current state is conditionally independent of the path of past states, a characteristic called the Markov property. In the current application the states of the Markov process correspond to crack penetration depths in the material, and the transition probabilities from a lower state to higher states (deeper cracks). The effects of inspections are included in the model as transitions from a crack state to a flawless state. A similar approach to study inspection strategies using Markov models has been suggested in [Ref. 8]. The main novelty here is the use of PFM modelling to generate transition probabilities to model the crack growth.

The principle of a simple Markov model is illustrated in Figure 12.

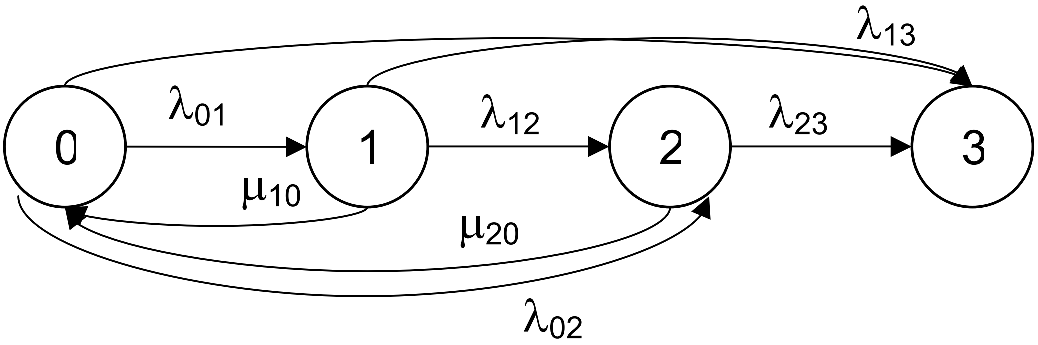


Figure 12 Illustration of a Markov model for degradation and repair

In this illustration we have defined the following states: 0 = no detectable flaw, 1 = a detectable flaw, 2 = detectable leak, 3 = rupture. λ_{ij} and μ_{ij} are transition rates (or probabilities per cycle) from state i to state j. λ_{ij} symbolise growh rates, while μ_{ij} describe detection and repair rates.

The state transition matrix for the above system is:

$$M = \begin{pmatrix} 1-\lambda_{01}-\lambda_{02}-\lambda_{03} & \lambda_{01} & \lambda_{02} & \lambda_{03} \\ \mu_{10} & 1-\mu_{10}-\lambda_{12}-\lambda_{13} & \lambda_{12} & \lambda_{13} \\ \mu_{20} & 0 & 1-\mu_{20}-\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{10}$$

In our discretised model, we use probabilities instead of rates. We can also define several states for different flaw depths, and define different flaw detection probabilities for each depth state.

Transition probabilities from one state to the other higher states are generated from the results of the PFM simulations. From the simulations we obtain information on the flaw depth once a year, and this size is assigned to a state representing a certain range of depth. The data tell how many years has been spent in each of these states, and if there is a transition to a higher state. The transition probability frpm state *i* to *j (j>i)*, p_{ij}, is obtained by dividing the number of transitions from *i* to j by the number of years spent in state *i*. In theory any number of states can be used, but the PFM simulation step limits this number. Depending on the application a suitable number of states is between 5 and 10.

The general discrete-time Markov process is:

$$\overline{p}_t = \overline{p}_{t-1} \times M \tag{11}$$

where $\overline{p}_t$ (*t* = 1,2,3…) is a vector containing the state probabilities (i.e. probability that the flaw has advanced to that state) for step t, and M is a matrix containing the state transition probabilities. $\overline{p}_0$ is a vector describing the initial flaw depth distribution of the component.

For better modularity and ease of application this basic Markov process is modified to use two different state transition matrices. The degradation matrix M_{deg} models the crack growth, and the inspection matrix M_{ins} models the inspections. While the degradation matrix affects the state probabilities each year, inspections affect it only on those years inspections are performed, yielding the following modified discrete-time Markov process:

$$\overline{p}_t = \overline{p}_{t-1} \times M_{\deg} \times \left(B(t) \cdot M_{ins} + (1 - B(t)) \cdot \mathrm{I}\right) \qquad (12)$$

where *B(t)* is a Boolean function with value 1 if inspections are performed at time step *t* and 0 if no inspections are performed at time step t, and I is a unit matrix. While any length can be chosen for the time step, one year seems natural for RI-ISI application, since the outages during which inspections could be performed are usually once a year.

The Markov process was coded and the results were calculated with Matlab software.

8.2 Examples of Simulations

We illustrate the development of crack growth and the effect of inspections with some examples. The examples cover different degradation mechanisms: stress corrosion cracking and thermal fatigue.

8.2.1 Stress corrosion cracking initiation and growth

Our first example assumes intergranular stress corrosion cracking initiation and growth in an austenitic piping weld. Here we also demonstrate how a decrease in POD can be compensated with a shortened inspection interval.

Let us assume two alternative POD step functions, as shown in Figure 13. In both cases the POD is 0 for cracks smaller than 20% of the wall thickness. For cracks exceeding 20% of the wall thickness, POD 1 is 0.9 and POD2 is 0.8.

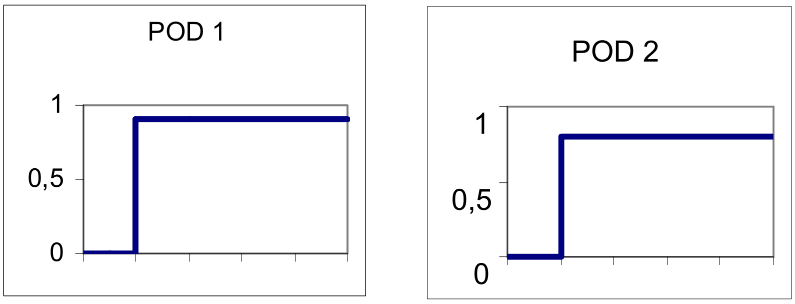


Figure 13 POD functions used in the example.

Figure 14 shows the yearly rupture probabilities for three cases: without inspections, inspections with POD1 at 10 years interval and inspections with POD2 at 8 years interval. In this case, where we have no manufacturing flaws, but the IGSCC cracks are initiated according to a frequency and size distribution estimated from Swedish experience data, and the annual rupture probability is an increasing function.

The mean yearly rupture probability for the case without inspections is 3.7E-5, for POD1 case it is 2.3E-6, and for POD2 case 2.4E-6. Thus a similar yearly rupture probability is achieved by shortening the inspection interval for the lower quality inspections. In terms of the risk reduction measure defined in Section 5, this is a risk reduction of 94 % calculated over the assumed 60 year component lifetime.

8.2.2 Crack growth due to thermal fatigue

In a second example, we illustrate a case with thermal fatigue. In the case of thermal fatigue induced cracking it is typical that a crack either grows very fast or hardly at all. Two PODs are considered: POD1 is the same as in the previous case. For the other (POD3), the POD for cracks exceeding 20% of the wall thickness is only 0.65.

Figure 15 shows the development of yearly rupture probabilities without inspections and the different inspection strategies, POD1 and POD3 both with 10 year inspection intervals. Yearly rupture probability for the two POD cases are almost identical despite POD3 having much lower probability of detection than POD1. Mean yearly rupture probability without inspections is 1.7E-4, for POD curve 1 it is 1.5E-4, and for POD curve 5 it is 1.5E-4. The risk reduction is only 12%. In a case like this it is unlikely that an inspection will reveal a crack before it grows through wall.

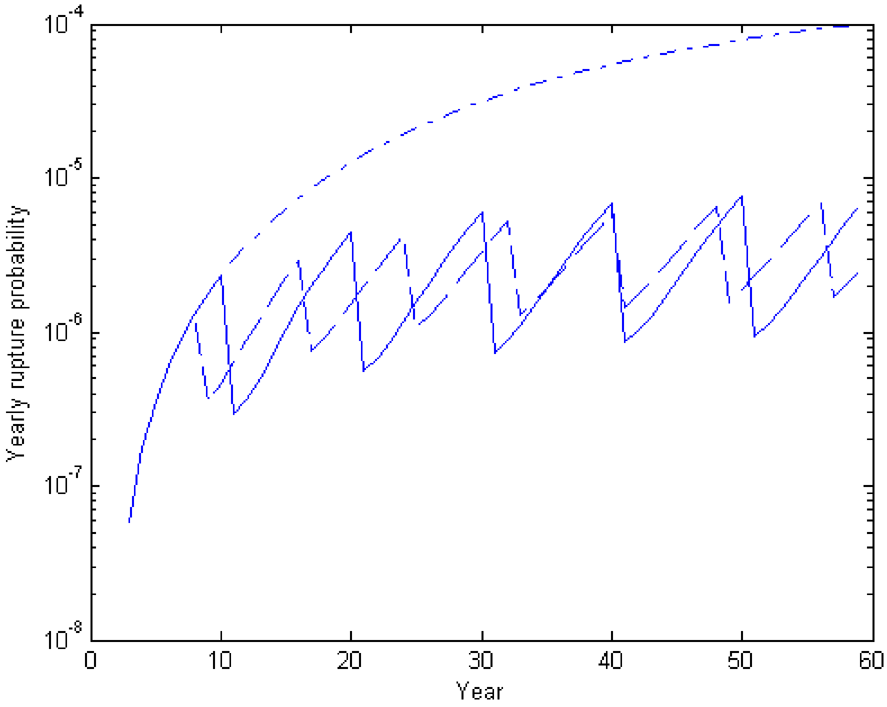


Figure 14 Yearly rupture probability for POD1 (solid line), POD2 (dashed line), and no inspections (dash-dotted line), IGSCC initiation case.

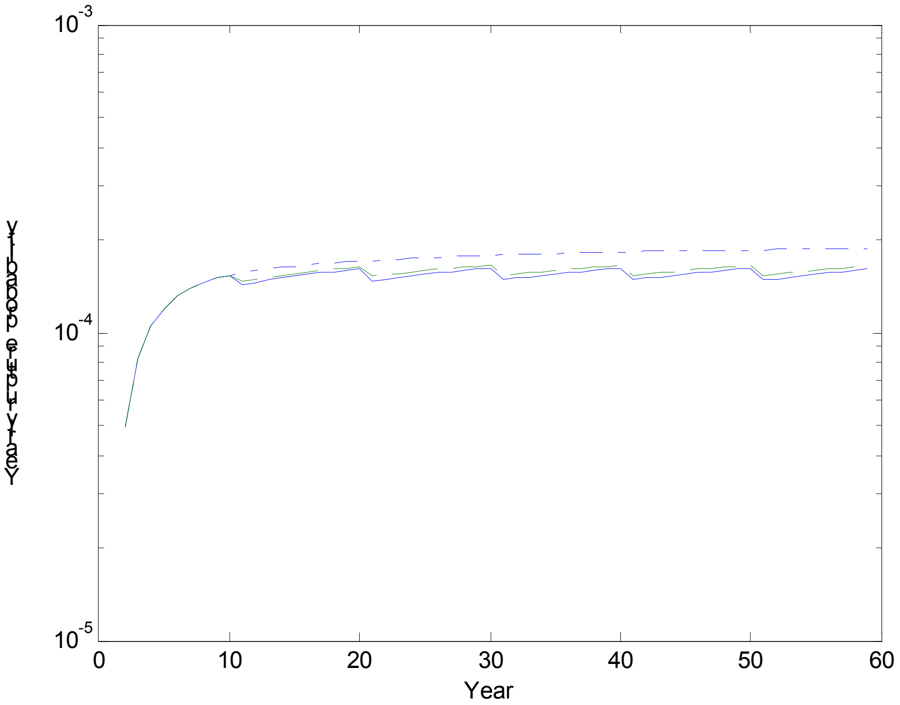


Figure 15 Yearly rupture probability for POD1 (solid line) and POD3 (dashed line), and no inspections (dash-dotted line), thermal fatigue case.

8.2.3 Concluding remarks

The examples of this section illustrate an approach to account for the flaw initiation and growth in assessing the effect of inspections on the failure probability of a structural component. Besides the POD level and interval assumptions, the assumptions on the defect growth have an important effect on the efficiency of the inspections. If the growth is very fast compared to possible inspection intervals, it is obvious that other means for reducing the failure probability should be applied.

These examples are based on a number of assumptions and limitations, related for instance to the randomised parameters of the probabilistic fracture mechanics model. The absolute values are thus subject to large uncertainties. The Markov property assumptions made for the transition from one flaw depth state to another can be questioned, since for some degradation mechanisms, it might be justified to assume a “memory” (the future growth depends not only on the present flaw size, but also on how this size has been achieved). Further, the assumptions in the inspection process are highly simplified, assuming that detecting a flaw returns the component to a flawless state. Despite these simplifying assumptions it is believed that the modelling approach gives reasonably realistic results for the comparison of alternative inspection strategies.

9 PILOT QUALIFICATION EXERCISES

9.1 Training of the Qualification Body members

A one-day meeting was organised with the purpose of training the members in the relevant fields not strictly falling within their expertise areas, such as probability theory, etc.

The day was structured around three main sessions. The first was dedicated to giving an introduction to expert judgement, with a focus on the biases and heuristics that are commonly encountered when eliciting probability statements from experts. The second was dedicated to revising the general concepts of probability relevant to the experimental determination of a population proportion. The third main session was dedicated to the Bayesian model for the quantification of the ENIQ methodology. In particular, worked examples were discussed. Finally, a discussion on the pilot study exercise (structure, duration, type of questions asked, etc.) took place.

9.2 1st Pilot Study

The purpose of the pilot qualification study was to study and as far as possible demonstrate approaches to quantifying the outcome from inspection qualification.

These approaches varied from simple expert judgement to the application of a Bayesian methodology. The lessons learnt from the pilot study were taken into account before finalising the guidelines provided on how to quantify the outcome from inspection qualification.

9.2.1 Preparation

It was intended to trial three approaches:

- Use of direct judgement to decide whether a simplified POD curve (e.g. a step function) can be considered as a lower bound for actual inspection capability (See Section 6)
- Use of an experimentally derived relationship between margin of detection (e.g. signal to noise, signal to report threshold) and POD (See Section 6)
- Use of the Bayesian method (See Section 7)

Note that these approaches can be used in combination, e.g. approach (b) could be used in combination with approach (a) or (c).

A technical justification (TJ) was provided for the pilot by a utility. This original TJ was modified for technical reasons and to respect the confidentiality of the utility which provided the dossier. Two variants of the TJ were produced with differing technical content in order to study the sensitivity of the results to TJ content. The TJs were referred to as TJ1 and TJ2.

The use of a genuine TJ as the basis for the modified TJs was intended to ensure as far as practicable the realism and credibility of the pilot.

9.2.2 Programme

The pilot qualification body (QB) comprised J Whittle (J Whittle & Associates) as chairman, R Booler (Serco Assurance), H Söderstrand (SQC) and B Dikstra (Doosan Babcock). All have extensive experience of inspection qualification.

L Gandossi, K Simola and B Shepherd had been involved in developing the quantification approaches and acted as facilitators. The QB was responsible for reviewing the TJs and applying (as far as the members felt able) the quantification methods.

The ultimate decision on whether the outcome from any of the TJ variants could be quantified and what the quantified value was, rested with the QB.

The pilot study was held in August 07 and lasted just under 3 days, including a short initial refresher session on the quantification methods and to address any queries before starting the quantification process. The programme was based on working through TJ2 using the three different approaches (a,b,c) in turn, and then repeating for TJ1.

The order in which the approaches were applied was to be a, b, c to avoid the results of the more rigorous approach(es) influencing those based on direct judgement.

9.2.3 Technical Justifications

Two variants of the original TJ were produced. The original intention was that one TJ would present more information than the other, and that the “less complete” one would be quantified first, before revealing the additional information in the second TJ. It was thought that this would lead to more objective assessments than performing quantification of the more complete TJ first, then trying to ignore some of that information.

It was subsequently decided that removing some of the information in the TJ could make the exercise unrealistic (artificially low quality of TJ) so instead it was decided to keep the same amount of information, but make it less convincing by reducing signal amplitudes in one of the TJs (see below for more detail).

TJ1 and TJ2 covered the same component which was a circumferential weld in a large diameter vessel. Part of the weld had been excavated to remove defects prior to repair welding. The TJ covered both manual and automated ultrasonic procedures, which were based on the same probes, scanning patterns and reporting thresholds.

The contents of each TJ included among other things input information (details of geometry, access, defects to be detected etc), physical reasoning to identify worst case defects, results of modelling to predict defect responses, and experimental evidence from open trials on test pieces.

TJ1 and TJ2 were identical except that

- all relevant experimentally measured and theoretically derived signal amplitudes from defects of concern were 4dB higher in TJ1 than in TJ2 (i.e. TJ1 represented improved inspection capability compared to TJ2);

- some of the text was modified to reflect these differences, e.g. a phrase like “indications were well above threshold” in TJ1 was changed to “indications were above threshold” in TJ2.

9.2.4 Quantification of TJ2

9.2.4.1 Direct Judgement

Each member of the QB had independently completed a table recording their direct expert judgement of POD separately for the manual and automated UT covered by TJ2. The principle was that they ticked boxes representing progressively higher POD values which they considered were below actual POD, until they reached a POD value where they had reservations (and placed a “?” against that value). The results were presented without any anonymity and were as follows (Table 3):

Manual Ultrasonic inspection

POD
Inspection exceeds

	HS	BD	JW	RB
50%	√	√	√	√
60%	√	√	√	√
70%	√	√	√	√
80%	?	√	√	√
90%		?	?	√
95%			?	?

Automated Ultrasonic inspection

POD
Inspection exceeds

	HS	BD	JW	RB
50%	√	√	√	√
60%	√	√	√	√
70%	√	√	√	√
80%	√	√	√	√
90%	?	√	√	√
95%		?	√	√

Table 3 Direct Judgement for TJ2 (without confidence levels)

Each member of the QB explained the reasoning behind their assessment. The QB members were then asked to quantify their confidence that the maximum POD values they had stated were lower bounds for inspection capability. They were as follows (Table 4):

QB Member	Manual	Automated
HS	95%	95%
BD	90%	95%
JW	80%	90%
RB	80%	90%

Table 4 Confidence levels for maximum PODs (TJ2)

Following this discussion the QB members were asked to state what lower bound POD they judged corresponded to a confidence level of 95%, and a confidence level of 80%. The results are presented in Table 5a and Table 5b.

Confidence	HS	BD	JW	RB
95%	70%	60%	60%	70%
80%	80%	90%	80%	90%

Table 5a Lower bound PODs for different confidence levels (TJ2 manual)

Confidence	HS	BD	JW	RB
95%	80%	80%	90%	90%
80%	90%	95%	95%	95%

Table 5b Lower bound PODs for different confidence levels (TJ2 auto)

It was decided not to try to achieve an overall QB consensus on the direct judgement POD values and confidences for TJ2. It could be left to RI-ISI analysis to decide e.g. whether to average values from QB members, or whether to adopt the most pessimistic values. It was noted that in any event there was reasonable consensus, with a lower bound POD of 85% ± 5% at 95% confidence level, for the automated UT procedure (lower bound POD 65% ± 5% at 95% confidence for manual).

9.2.4.2 Use of Relationship between POD and Margin of Detection

For ultrasonic inspection, as defect response increases above the reporting or noise level, it is to be expected that the probability of detecting and reporting defects will tend to increase. Understanding the relationship between this margin of detectability and the probability of detection / reporting could be a useful tool during inspection qualification, and could help to quantify the outcome.

SKI have funded preliminary work by Doosan Babcock to investigate this relationship. A report on the work has been published on the SKI website [Ref. 5].

A document “Relationship between Probability of Reporting and Signal above Report Threshold” drawing on the results of this work was distributed to the QB members. A figure in the document illustrated a partly hypothetical graph relating the probability of an operator judging that a defect was above the report threshold, as a function of how high the signal actually was above the report threshold.

According to TJ2, the signals from defects of concern should be significantly above the report threshold. The probability of reporting a detected defect was therefore high (above 95% according to the graph). If defects were not reported, it was therefore unlikely to be due to variations in setting report threshold or measuring defect signal amplitude. The reasons were more likely to be due to other human factors such as deviations from the specified scan pattern, couplant loss etc (as well as worst case combinations of defect size, roughness, skew etc.)

It was therefore agreed that for this specific TJ, the relationship between probability of reporting and signal above threshold was not particularly useful.

9.2.4.3 Use of Bayesian Methodology

There was some general discussion before starting to apply the Bayesian Methodology.

It was agreed that it would only be applied to the automated UT procedure. It was considered that determining the relative weights of the TJ (excluding trials information) and the trials information was a difficult step. It was also agreed that it was important to avoid ending up with an over-pessimistic value for POD due to taking into account very unlikely combinations of worst-case characteristics. It was difficult to know how to address the reduced detectability of worst case defects without knowledge of what proportion of the overall defect population they represented. For example if the requirement is to detect defects with skew between -5° and + 5°, then it might be reasonable to assume a Gaussian distribution within these limits. However this would mean that the proportion of defects at the limits of skew was almost zero, even though TJs often concentrate on worst case defects.

The application of the Bayesian Methodology then started with the QB members deciding what the elements of the TJ were (excluding experimental evidence which was to be considered separately). After extensive discussion the elements of the TJ which were to be individually weighted and scored were selected as:

a) Physical reasoning for non-worst case defects
b) Modelling and physical reasoning relating to worst case defects
c) Personnel
d) Equipment

It was agreed that the relative weights between a) and b) should reflect the relative proportions between non-worst case and worst case defects.

Each QB member then stated what weight and score he gave to each of these TJ elements. These were as shown in Table 6a and Table 6b.

TJ element	HS	BD	JW	RB
a	40	40	40	50
b	30	10	20	30
c	20	25	20	10
d	10	25	20	10

Table 6a Weights for TJ2 elements

TJ element	HS	BD	JW	RB
a	1	1	1	1
b	0.95	0.6	0.7	0.95
c	0.95	0.99	0.95	0.95
d	1	0.90	0.90	0.95

Table 6b Scores for TJ2 elements

The discussion moved on to how to weight and score the trials results. Two views emerged regarding what to count as the number of “hits.” One view was to take account of the number of times a defect was detected by different beams so that a defect detected by 3 different beams would count as separate 3 hits. The other view was to ignore this diversity of detection and only consider whether a defect had been detected overall, or completely missed. It was decided that for this current exercise, diversity of detection would be ignored (i.e. simply count number of defects detected) but it was recognised that this was an issue which would require further consideration in any future guidelines which were produced..

It was also agreed to consider any defect which had been detected with signal above record threshold as a “hit” without taking into account any margin by which the signal was above this threshold.

During the weighting of the elements of the TJ, the average weighting ratio of the element relating to non-worst case defects (element a) to worst case (element b) was approximately 2.5. Since the trials presented results from 6 defects, the TJ could be considered equivalent to (2.5 x 6 non worst case defects) plus (6 worst case defects) = 21 defects.

Lower bound POD for a given confidence level was then determined using a spreadsheet which performed the calculation according to the Bayesian methodology. The data entered into the spreadsheet were the TJ element scores and weights, the equivalence of the TJ to 21 defects in test piece trials, and 6 hits out of 6 in actual experimental.

The results are presented in Table 7.

Confidence	HS	BD	JW	RB
95%	87%	82%	80%	87%
80%	92%	88%	86%	92%

Table 7 POD for TJ2 using Bayesian Methodology

It was noted that there was reasonable consistency between the PODs from direct judgement (80% - 90% at 95% confidence) and that derived using the Bayesian Methodology (80% - 87% at 95% confidence).

9.2.5 Quantification of TJ1

9.2.5.1 Direct Judgement

The QB members were provided with a copy of TJ1 together with a note which highlighted the differences between TJ1 and TJ2. The key difference was that all relevant experimentally measured and theoretically derived signal amplitudes from defects of concern were 4dB higher in TJ1 than TJ2, representing improved inspection capability.

The QB members then each stated what they considered a lower bound POD to be, for 95% and 80% confidence levels, based on direct judgement. The results are presented in Table 8a and Table 8b.

Confidence	HS	BD	JW	RB
95%	80%	75%	70%	85%
80%	90%	95%	90%	90%

Table 8a Lower bound PODs for different confidence levels (TJ1 manual)

Confidence	HS	BD	JW	RB
95%	90%	90%	90%	90%
80%	95%	99%	95%	95%→97.5%

Table 8b Lower bound PODs for different confidence levels (TJ1 auto)

9.2.5.2 Use of Relationship between POD and Margin of Detection

It was agreed not to apply the relationship between probability of reporting and margin of detection to TJ1, for the reasons explained in 3.1.4.2

9.2.5.3 Use of Bayesian Methodology

As with TJ2, it was decided only to apply the Bayesian approach to automated UT. It was agreed that the weights of the elements in TJ1 would not differ from those for TJ2, nor would the relative weights of the TJ and the experimental trials change.

The weights and scores for the TJ1 elements are presented in Table 9a and Table 9b.

TJ element	HS	BD	JW	RB
a	40	40	40	50
b	30	10	20	30
c	20	25	20	10
d	10	25	20	10

Table 9a Weights for TJ1 elements

TJ element	HS	BD	JW	RB
a	1	1	1	1
b	0.95	0.9	0.9	0.97
c	0.97	0.99	0.95	0.95
d	1	0.90	0.9	0.95

Table 9b Scores for TJ1 elements

Lower bound POD for a given confidence level was then determined using a spreadsheet which performed the calculation according to the Bayesian methodology. The data entered into the spreadsheet were the TJ element scores and weights, the equivalence of the TJ to 21 defects in test piece trials, and 6 hits out of 6 in actual experimental.

The results are presented in Table 10.

Confidence	HS	BD	JW	RB
95%	87%	85%	84%	87.5%
80%	92.5%	91%	89.5%	92.5%

Table 10 POD for TJ1 using Bayesian Methodology

As with TJ2, there was reasonable consistency between the PODs from direct judgement (90% at 95% confidence) and that derived using the Bayesian Methodology (84% - 87.5% at 95% confidence).

9.2.6 Debriefing Discussions

After completing the quantification processes for the two TJs, a general discussion was held. It was suggested that a more appropriate method of applying the quantification methodologies might be to consider different categories of defect separately, and to determine POD separately for each.

The QB would then state what the POD associated with each category of defect was (which could cover non-worst case defects, and different types of worst case defects) and it would be the responsibility of those performing the structural integrity analysis to make assumptions about the relative populations of these different categories.

TJs could be constructed to facilitate the quantification process, and this could lead to further useful work within ENIQ (guidelines on how to construct such a TJ). It was recognised that the method of determining the relative weights of the TJ and trials when applying the Bayesian approach was not robust – the QB could have come up with a different method, or split the TJ into different elements

It was considered in retrospect that it would have been better to split the TJ into different elements from those chosen. It was agreed that a suitable split (which could be generic) was:

- Equipment
- Personnel
- Coverage
- Detection margins
- Evaluation
- Independent evidence

Overall, the pilot exercise was considered to have been a success even though further work was required to establish robust guidelines.

9.3 2nd Pilot Study

The purpose of the second pilot qualification study was the same as for the first pilot study (JRC Petten, August 2007), i.e. to study and as far as possible demonstrate approaches to quantifying the outcome from inspection qualification. Any appropriate lessons from the first pilot were to be taken into account.

9.3.1 Preparation

Two approaches were trialled:

- Use of direct judgement to decide whether a simplified POD curve (e.g. a step function) can be considered as a lower bound for actual inspection capability.

- Application of the Bayesian method.

A technical justification (TJ3) was provided for the pilot by a Swedish organisation. This TJ was not from a real application but had been produced as an example of the scope and content of a TJ. This original TJ was edited in case there was a need to respect confidentiality of the organisation which provided the dossier.

The following were issued to the members of the pilot qualification body before they convened for the study:

- Technical Justification version TJ3
- Draft Programme
- Direct Judgement form for TJ3 to be independently completed and returned by the QB members before they convened.

9.3.2 Programme

The terms of reference, strategy for the pilot and roles of the participants were as for the first pilot study. The same participants were involved:

The members of the qualification body were J Whittle as chairman (J Whittle & Associates), R Booler (Serco Assurance), H Söderstrand (SQC), B Dikstra (Doosan Babcock).

The facilitators were L Gandossi (JRC), K Simola (VTT) and B Shepherd (Doosan Babcock).

The pilot study lasted just under 2 days, including a short session to address any queries before starting the quantification process.

9.3.3 Technical Justification and Preliminary Discussions

TJ3 covered automated UT of a 36mm thick circumferential butt weld in a large diameter ferritic vessel.

The contents of the TJ included among other things details of geometry, access, defects to be detected, physical reasoning to identify worst case defects, results of modelling to predict defect responses, and experimental evidence from trials on test pieces.

It was noted that the TJ did not state anywhere that there was no requirement to detect lack of fusion. It was agreed that for the purpose of this study only the defects explicitly stated as requiring detection would be considered, although this highlighted the importance of clearly stating inspection objectives.

It was also noted that there was an implicit assumption within the TJ that defects larger than the target size would be easier to detect but this was not always necessarily the case. It was commented that it was common to all TJs that the worst case defect be correctly determined and that when determining POD it is important to establish what the population covered.

There was some discussion on whether for this study POD should mean probability of both detecting and correctly characterising a defect (since characterisation played a key role in TJ3). It was agreed to start off by quantifying only detection probability, and to address characterisation subsequently if there was enough time.

One of the QB members stated that an issue he disliked in TJ3 was that the sizing tolerance was larger than the target defect size – in Sweden normal practice is to subtract the tolerance from the size which must be detected in order to establish the qualification size. However in this instance it was not important since the TJ was only an example.

9.3.4 Quantification of TJ3

9.3.4.1 Direct Judgement

Each member of the QB had independently completed the direct judgement table. The TJ gave 18mm x 90mm as the qualification size. However virtually all the TJ data refers to the target size of 4mm x 20mm and so this was the basis of the POD assessment.

The results were presented without any anonymity. The results are presented in Table 11 below.

Confidence	HS	BD	JW	RB
95%	80%	80%	90%	90%
80%	95%	90%	95%	95%

Table 11 Lower bound PODs for different confidence levels (TJ3)

The members of the QB then discussed their judgements.

JW considered that the signal response should always be above the record threshold by a significant margin so that causes of failure would be mainly due to human factors in the analysis process. He was therefore 95% confident that POD would be at least 90%.

RB stated that for unskewed surface breaking smooth cracks the POD should be high but the cracks covered by this TJ could be rough and could be skewed, and there was not much detail in the TJ addressing these aspects. Surface form and coupling variations could also influence detection performance. He was therefore 90% confident that the POD was at least 90%, but was prepared to be more optimistic for the lower confidence of 80%.

The main reason BD judged the POD to be lower than the above was because of the cavities in the surface caused by grinding. He considered the treatment in the TJ to be optimistic since it did not address loss of coupling or reduction in signal amplitude. There was no information on how widespread the cavities were. Also the tilt and skew ranges covered were quite large and the effects of these would become more significant with increasing defect size. Since only one probe was employed human factors would also be more significant.

HS held similar views to BD, and considered that the combination of cavity and defect tilt could significantly reduce performance but overall was reasonably satisfied with potential inspection capability.

JW commented that in a real case the QB would ask the utility for more information on surface condition.

RB stated that he still felt uncomfortable concluding that 1 in 10 defects could be missed. BS replied that the POD score represented a minimum POD, e.g. reporting POD to be at least 90% (in the direct judgement table issued) could mean POD was considered to be between 90% and 95%.

There seemed generally to be more psychological reluctance to judge POD as being above 95%, or to quantify it more precisely than ± 5%, than to judge that fewer than 5% of defects would be missed, or to quantify the percentage of defects missed more precisely than ±5%.

RB thought that more knowledge of the relative distributions of different classes of defects would result in more consistency between probability of detecting, and probability of not detecting, judgements.

BS asked whether the approach adopted in the direct judgement table, whereby progressively higher POD's are systematically ticked, helped to overcome initial reluctance to assign a POD to qualification results. HS confirmed it was useful.

In response to a query from JW, BD replied that if the surface was smooth, he would judge the POD to be closer to 98% - 99%.

HS said he viewed things a bit differently. During qualification, represented inspection conditions were generally good whereas a number of factors (environment, access, set-up etc) could reduce inspection effectiveness. He therefore took these issues into consideration when judging POD. BD agreed that the confidence from inspection qualification should not be oversold. However RB was of the opinion that all these potential negative factors should be taken into consideration during qualification.

It was agreed that (as with the 1st pilot study) use of the relationship between POD and margin of detection was not appropriate for this TJ.

9.3.4.2 Use of Bayesian Methodology

It was agreed that a useful starting point for application of the Bayesian method would be to adopt the TJ elements established at the end of the first pilot. It was recognised that any future guidance on element selection should be applicable to all TJs.

JW commented that since the trials in TJ3 were an integral part of the justification it would be difficult to consider them separately without destroying the logic of the rest of the TJ

RB stated that he felt uncomfortable about adopting too rigorous an approach to weighting the TJ and trials – perhaps there was too much reluctance to rely on expert judgement.

The elements of the TJ were agreed to be:

- Equipment
- Personnel
- Coverage
- Detection margin
- Evaluation
- On site application

These were the proposed “generic elements” established during the 1st pilot debrief except that for element f “on site application” replaced “independent evidence”. (It was not clear what had been meant by “independent evidence” but RB thought it might influence other elements such as choice of probes, so not independent.)

There was a general discussion on the approach to determining weights, scores, and equivalent number of practical trials. Each member of the QB then stated what they judged the weight/score for each of the above elements to be. The results are presented in Table 12 below.

TJ element	HS	BD	JW	RB
a	20/1	20/1	15/1	10/1
b	10/0.9	10/1	15/0.9	5/1
c	20/0.9	20/0.8	15/0.7	15/0.9
d	35/1	20/0.75	40/1	60/1
e	5/0.95	20/0.9	5/0.7	5/0.8
f	10/0.95	10/0.7	10/0.7	5/1

Table 12 Weights and scores for TJ3 elements

The reasons for the above weights/scores were then discussed (In some cases QB members revised their judgements during the discussion. Table 12 presents the final figures).

RB weighted element d (detection margin) highly since he considered it the most significant issue in the TJ – as long as there was good coverage other issues shouldn’t be particularly important. Neither modelling nor trials included rough defects. LG confirmed to RB that in that case weight is reduced.

BD stated he found himself getting confused over

1. volume of evidence
2. how convincing the evidence was
3. relative importance of the issue
4. impact on inspection

This prompted further discussion on how to determine weights and scores. JW commented that if information is missing in an element, this would reduce its weight but this would result in other elements being weighted more highly. RB stated that if e.g. information is missing on rough defects, this doesn’t mean that the score should be reduced since there is no evidence that rough defects would be missed.

BD stated that this was why he was of the opinion that weights should not add up to 100%, but to some figure which represented the total amount of supporting information within the TJ.

The QB members considered it difficult to decide how many equivalent trials the TJ was worth. The sensitivity of the probability distribution curve to different equivalent trials figures was studied as a means of selecting a figure which appeared to give sensible results. BD expressed concern over selecting a figure which appeared to give the desired answer, although RB stated that there didn’t appear to be an alternative.

HS commented that he found the Bayesian approach a useful statistical tool for helping to decide the relative balance between a TJ and trials in order to achieve a specific level of confidence in a specific POD.

RB felt that further guidance was required on how to determine the equivalent number of trials. Although he supported the development of these quantification approaches it was important that they be robust enough to convince those in the qualification industry of their potential value.

KS stated that there still appeared to be confusion over what was meant by weight and score. She thought that it would be appropriate to have some method of comparing the weight of an element to its “ideal”.

At the start of the 2nd day of the pilot study, BD gave a short presentation he had prepared, prompted by some of the problematic issues which had arisen on equating the TJ to a number of equivalent trials.

The approach which BD described was based on considering the probability that a defect claimed to be detectable would actually be detected. A quantitative estimate could be made by looking at the detection margin between the predicted signal amplitude and the reporting threshold. This detection margin is in practice reduced by uncertainty in the setting of the threshold.

Since this uncertainty should normally be quantified in the TJ it should be possible to estimate the probability of failing to detect a given defect as a result of the threshold uncertainty exceeding the detection margin. The reciprocal of this probability would give an estimate of the number of trials for one detection failure. This could then be the basis of the number of trials equivalent to the TJ in the Bayesian analysis.

The analysis could take account of uncertainty in the reporting process as well as in the setting of the threshold. In the case of manual UT, this has already been studied experimentally (SKI funded work on relation between probability of detecting / reporting and margin of detection) so that quantitative information is available. For automated UT, a simple allowance might be made for the margin necessary for the analyst always to register the presence of a reportable indication.

RB pointed out that the reporting criteria for automated UT may often include factors other than amplitude (e.g. the extent of the plotted UT indication in one or more dimensions) so that a quantitative analysis might be difficult.

In further discussion it was recognised that arguments of the type presented might alternatively provide a means of scoring the TJ (or some of its elements). This line of argument could not however be used to both score the TJ and weight it relative to experimental trials.

The QB members felt that there was still confusion over what was meant by weights / scores, and the differences in these figures from the different members was probably due to different interpretations of these factors rather than significant differences of opinion on the effectiveness of the inspection embodied in the TJ. Strong guidance was needed.

RB suggested that it could be useful to validate the methodology and guidance by performing the Bayesian analysis on a simple procedure / TJ then performing practical trials to measure POD.

JW stated that there was also a need to establish how to deal with missing information since different views had emerged on this.

KS replied that she liked the concept of establishing what weights the TJ would have if it was perfect, and then marking the weights down where information was missing. HS commented that there might be a good chance of getting further funding to develop and validate more detailed guidance.

BS commented that if an individual element indicated a limitation in inspection capability, e.g. a limitation due to incomplete coverage, then this should represent a lower bound for overall capability. Thus if coverage was 75% then this should be an upper bound for POD to reflect this weak link in the chain. With the current methodology this limitation would appear to be "diluted" by the other elements – possibly there should be some means of multiplying (or convoluting) factors representing the score of the individual elements.

JW replied that shortcomings in inspection capability tended to be due to either "massive" errors (blunders) or statistical variations. KS stated that it had been assumed that the elements were truly independent.

RB felt that the biggest conceptual leap was equating the TJ to a number of equivalent practical trials, but LG pointed out that this was a fundamental requirement of the methodology. BS stated that it might help to consider how much evidence from trials would be required, in order for the trials to be an adequate substitute for the TJ.

RB and HS said that even though there remained issues to be resolved before applying the methodology for real qualifications, they had already found the quantification concepts being studied under this project useful in stimulating more critical consideration of TJs and implicit assumptions in them.

Each of the QB members then stated how many equivalent trials they considered TJ3 to be worth. The results are presented in Table 13.

	HS	BD	JW	RB
Equivalent Trials	65	100	25	15

Table 13 TJ3 equivalent trials

The QB members then discussed their views (in some cases QB members revised their judgements during the discussion. Table 13 above shows the final figures).

BD had approached this from an order of magnitude perspective since it was difficult to be precise. He considered a good TJ (which this was) to certainly be worth more than 10 trials but did not consider TJ3 to be worth more than 100.

HS had imagined how he would design trials. He would want to include defects at the two extremities of tilt, the two extremities of skew, would want to include the effects of cavities etc. and then would want around 6 or 7 defects for each category.

JW felt that the evidence in the TJ which provides the most useful information was from the 14 defects for which results are provided, so on this basis the TJ was worth 14 which was then increased to 25 to allow for the fact that the defect designs were influenced by information in the rest of the TJ.

He asked BD whether, according to this logic, he thought that the extra information in the TJ was equivalent to trials on around 85 defects (100 minus the actual trials reported). BD confirmed that he thought so, bearing in mind he was adopting an order of magnitude approach. JW stated that the TJ addresses the limits of e.g. tilt and skew without taking into consideration whether these are worst cases.

RB had considered how many trials and successful detections would be required to generate the same POD and associated confidence as the TJ was judged to. To achieve a 95% confidence that POD was 95% would require around 50 defects. However there was a lack of information in the TJ, for example insufficient treatment of rough defects, so he considered the TJ to be equivalent to 15 defects, all detected.

He said that in BD's case, 100 trials would provide high confidence in POD, and further information would not be required, but this was not the case (since there was a lack of information on rough defects).

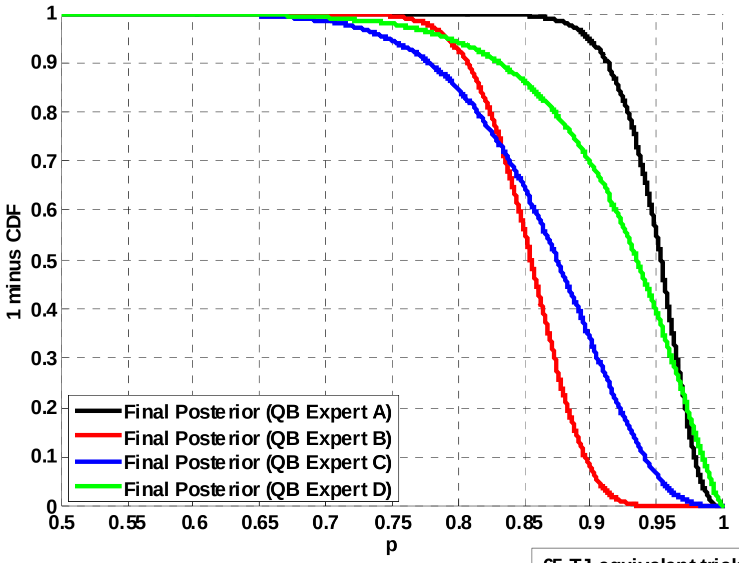
This prompted discussion which concluded that the equivalent number of trials should be the equivalent number of defects, not just the number of successful detections. The weights, scores and equivalent trials figures for the QB members were then entered into a MATLAB programme which LG had prepared to calculate POD as a function of confidence level. The results are presented in Figure 16 and Table 14. After the meeting, LG also provided results based on a figure of 25 equivalent trials for all QB members. These are presented in Figure 17 and Table 15.

Confidence	HS	BD	JW	RB
95%	90%	79%	75%	78%
80%	93%	82%	81%	87%

Table 14 Bayesian analysis (different equivalent trials)

Confidence	HS	BD	JW	RB
95%	83%	70%	75%	85%
80%	89%	77%	81%	91%

Table 15 Bayesian analysis (25 equivalent trials)



65 TJ equivalent trials for A (HS)

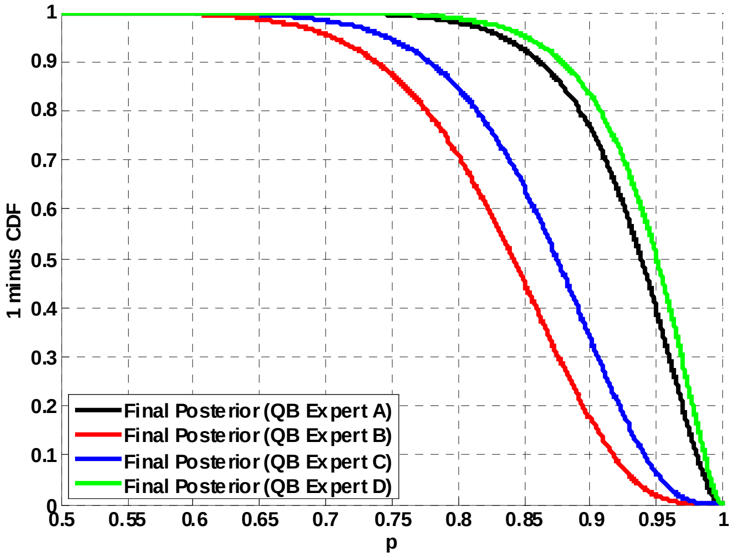
100 TJ equivalent trials for B (BD)

25 TJ equivalent trials for C (JW)

15 TJ equivalent trials for D (RB)

+ ZERO practical trials for ALL

Figure 16 Bayesian quantification results for TJ3 (different equivalent trials)



25 TJ equivalent trials for ALL experts and ZERO practical trials

Figure 17 Bayesian quantification results for TJ3 (25 equivalent trials).

9.3.5 General Discussion

After completing the quantification processes for TJ3, a general discussion was held.

RB pointed out that the four members of the QB had come up with four different POD assessments and asked how expert elicitation could be used to come up with a consensus. KS replied that although one approach would be for QB members to work together to agree figures, the simplest way would be to use an average.

BS drew attention to the previous comment made, that the differences were possibly due to differences in interpreting how to weight / score / decide equivalent trials rather than fundamental differences in opinion on the effectiveness of the inspection. It was agreed that more detailed guidance was needed.

RB stated that it would probably be better for QB members to work together to establish figures, rather than decide figures independently (as done in these pilots) since there could be an inclination to defend ones own figures. He thought that a follow on pilot based on this approach could be very useful, since the combination of the methodologies applied during the first two pilots, with expect elicitation, could be powerful.

HS stated that in Sweden that close teamwork was adopted whenever possible, and agreement is reached before qualification on which expert should concentrate on what.

10 EXAMPLE OF THE OVERAL PROCESS

Previous sections have described how a simplified POD curve might be used to represent inspection capability (Section 5), how this simplified POD curve could be used as the target (or output) from inspection qualification (Sections 6 and 7), and how POD, risk reduction and inspection interval can be linked (Section 8).

This section provides an example of the overall process which could be used to link RI-ISI and inspection qualification. The example is for illustrative purposes only, and no further conclusion should be made based on the numerical results in this example. Results depend on assumptions made on the structural properties and loads, assumed degradation mechanism, and on the number and definitions of the states in the Markov model (see Section 8.1).

In this example the components to be inspected are circumferential butt welds in ferritic steel. The welds are 36mm thick. The loading conditions and material properties are known and there is a postulated defect initiation and growth mechanism. From metallurgical considerations, a description of these potential defects can be provided (possible locations, orientations, roughness etc). The dimensions of the component and the growth mechanism correspond to the information available for the 2nd pilot study, summarised in Section 9.2.

Based on initial (non rigorous) opinions of NDT personnel, a reasonable target for inspection capability might be detection of potential defects exceeding 4mm throughwall extent and 20mm length with a POD of 95%.

The first step is to model defect initiation and growth, and to calculate the yearly rupture probabilities with and without inspection. This can be done using the approach described in Section 8. An inspection interval of 10 years is initially assumed.

The development of the yearly rupture probability is shown in Figure 18 both in the case of no inspections and in the case where the component is inspected at 10 year intervals with a 95% POD. The analysis indicates that with the inspection interval of 10 years the average failure probability (over the plant life) is reduced to 6.8E-7 from 6.0E-5 in a situation without any inspections. In terms of the risk reduction measure determined in Section 5, this corresponds to 98.9. In this example, we set the inspection requirement at about this level, i.e. the failure probability should be reduced by about 99%. Note that this is a requirement where the entire lifetime of the component is considered, and it implies several successive inspections.

The average risk reduction is not the only possible way to define the inspection target. Alternatively, the requirement could be set as a fixed average rupture frequency target, or even the maximum allowed yearly rupture probability (i.e. the peak values of the dotted line in Fig. 18).

The inspection requirements are then provided to the organisation which will perform the inspection: normally either the utility's internal NDT department, or a potential inspection vendor. These requirements state that automated ultrasonic inspection is required in order to maximise reliability. Having developed or identified the proposed inspection system (procedure + equipment + personnel), it is then subject to inspection qualification.

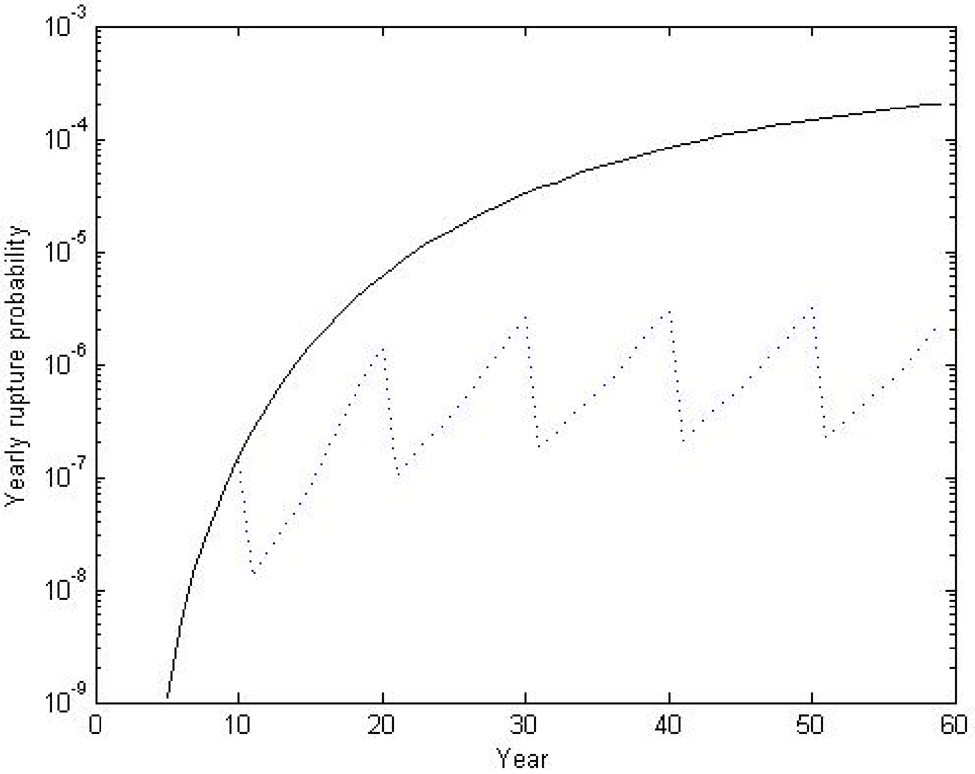


Figure 18 **Yearly rupture probability with no inspections (solid line) and with POD = 0.95 at 10 year inspection interval (dotted line).**

The inspection qualification body is given the task of determining whether defects which meet the defect description (location, orientation, roughness, etc) and which exceed 4mm through wall dimension and 20mm length can be detected with a POD of 95%. In order to provide improved transparency of the qualification process, the qualification body is instructed to employ the Bayesian methodology instead of direct judgement. The associated confidence level is also requested.

The inspection qualification body then attempts to quantify inspection capability, using one or more of the approaches described in Sections 6 and 7. In this example, the task is the same as one which was set in the 2nd pilot qualification study described in Section 9.1

We can therefore consider that a realistic outcome would be that the qualification body concludes that the POD 95% is not achievable by the proposed inspection system. The qualification body concludes that there is 95% confidence that the POD for the specified defects is at least 75% (see Section 9.2)

As the initially expected 95% POD could not be reached, the inspection interval should be adjusted to meet the initial criterion of reaching an average risk reduction of 99%. The analysis shows that given the POD of 0.75, an inspection interval of 5 years can be accepted. Also the 6 year interval results in a failure frequency very close to the criterion. Table 16 summarises the results for three different PODs and several inspection intervals. The red cell alternatives with risk reduction below 98.5% could be considered unacceptable.

Interval (y)	POD = 90%		POD = 80%		POD = 75%	
	rupt. /y	R	rupt. /y	R	rupt. /y	R
5			3.2E-7	99.5	5.0E-7	99.2
6			5.9E-7	99.0	9.0E-7	98.5
7	3.6E-7	99.4	9.6E-7	98.4	1.4E-6	97.7
8	5.7E-7	99.1	1.4E-6	97.7		
9	8.4E-7	98.6				

Table 16 Average rupture frequency (1/y) and the risk reduction percentage (R) for different PODs and inspection intervals

Figure 19 shows the yearly rupture probability for inspection interval 5 years and POD = 0.75. Also the case of no inspections is shown.

It is much easier to qualify against a simple step function POD, than to a more complex function. In Section 5 the sensitivity of the failure probability to the POD as a function of a set of assumptions (flaw size distribution, failure probability given a flaw size, form of POD) was illustrated. Here we study if, given the assumption made for the probabilistic fracture mechanic analysis in this example, providing more information on the detection capability of small defects would have a notable effect on the failure probability.

Instead of assuming no detection of any flaw smaller than 4 mm deep, it is assumed that defects between 2mm and 4mm are detected with 50% probability. In the case of POD = 75% for defects deeper than 4mm and 6 y interval, the risk reduction changes from 98.5% to 90.0%, and in the case of 7 y interval from 97.7 to 98.3. Even if the risk reduction improves slightly, this would not have a significant impact on the acceptable inspection interval.

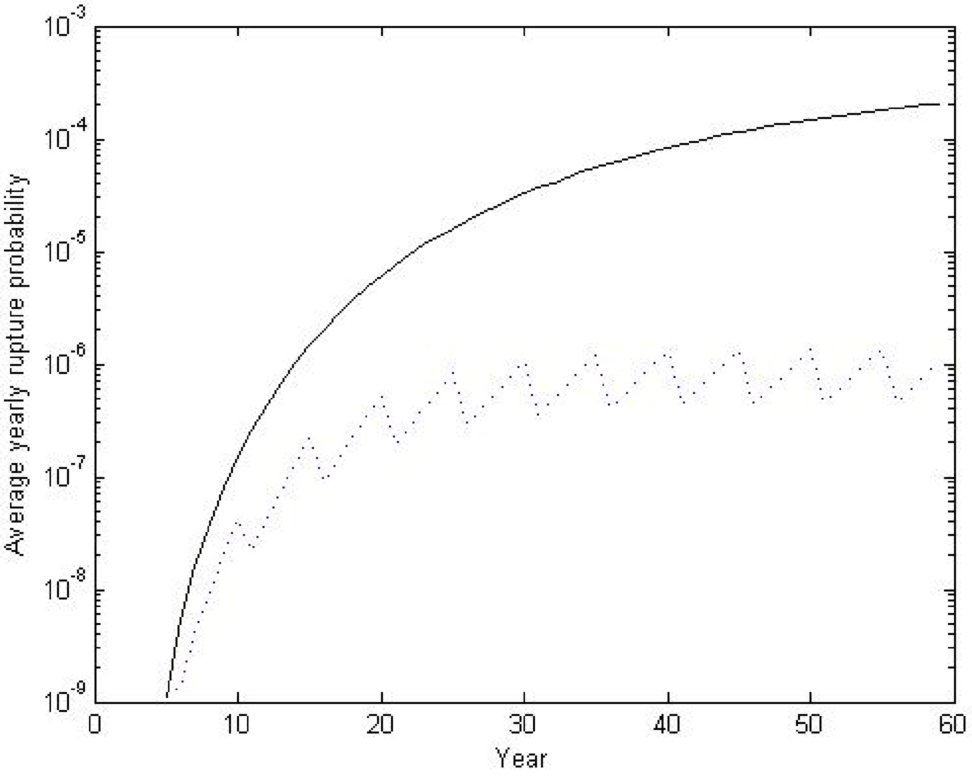


Figure 19 Yearly rupture probability with no inspections (solid line) and with POD = 0.75 at 5 year inspection interval (dotted line).

As a last point, let us assume that the main reason for the relatively low 75% POD according to the QB would be the detectability of the small defects close to the detection target. The QB would be confident that if the target detection size was set to 12mm, the POD would be the initially expected 95%. In this case, assuming that defects smaller than 12mm are not detected, the risk reduction target is achieved with a 6 y inspection interval.

11 CONCLUSIONS

The four main objectives of the project (see Section 3) were achieved. Each is reviewed in turn below.

11.1 Investigate Quantification Approaches

Approaches to quantifying the outcome from inspection qualification have been investigated and are described in Sections 5, 6 and 7. These are based on the use of a simplified POD curve (typically a step curve) to represent inspection effectiveness.

The work described in Section 5 has allowed sensitivity of risk reduction to the details of the POD curve to be investigated. Although the model which was the basis for this work is relatively simple, it provides a useful insight into the conditions under which a step POD curve could be adopted. It also illustrates the extent to which risk is sensitive to the plateau POD level in the step curve, and the defect size at which that plateau is reached (equivalent to defect qualification size). This approach can help achieve an appropriate balance between risk reduction and the effort involved in designing and qualifying an inspection system.

It should be noted that this work has studied relative risk reduction, i.e. the factor by which risk is reduced by inspection. A small relative risk reduction in a high risk component could be more beneficial than a high relative risk reduction in a low risk component.

The results of this work indicated that relative risk reduction was least sensitive to the detail of the POD curve for tough materials containing only relatively small defects (typical of nuclear applications).

Inspection quantification methods were developed based on direct judgement (Section 6), and a Bayesian approach (Section 7).

Direct judgement would involve the qualification body deciding whether, on the basis of all the information provided, they considered that a target POD curve represents a lower bound for actual inspection capability. Alternatively the qualification body could be asked to define a lower bound for the POD curve. In either case the task of the qualification body will be more difficult the greater the detail in the POD curve, the higher the maximum POD in the curve, and the lower the defect size (qualification size) corresponding the maximum POD.

A useful tool in making this judgement could be the relationship between POD and margin of detection (signal to report threshold, or signal to noise level) as described in Section 6.

The Bayesian approach is based on reviewing the Technical Justification and expressing the resulting (qualitative) degree of belief in inspection capability in probabilistic terms, by equating it to an equivalent number of trials and successes. This is then combined with the number of trials and successes from practical trials, after judging the relative weight of information from the TJ and from the trials.

11.2 Produce Guidelines

Practical guidelines on the application of these quantification methods were produced. The guidelines for direct judgement are relatively straightforward and are essentially the whole of Section 6. The Bayesian approach is more complex. While the technical approach is explained in Section 7 (and in more detail in Ref 6), separate guidelines which were finalised after the pilot studies are provided in Annex 1.

Section 8 illustrates how it can be possible to study the relation between the outcome from the qualification process (as expressed by a POD), risk reduction and inspection interval. Note that this requires knowledge of defect initiation and growth. The examples are based on a model developed at VTT combining probabilistic crack growth analysis with a discrete Markov process for inspection modelling. Alternative suitable probabilistic fracture mechanics tools could be used as well, provided they allow the modelling of inspection reliability. Section 10 provides an example of linking POD, risk reduction and inspection interval based on the 2nd pilot study.

11.3 Pilot Studies

The two pilot qualification studies were performed and were very useful in demonstrating the application of the quantification methods in practice, identifying issues which should be addressed in the final guidelines, and highlighting areas where further work was required. It should be noted that both pilots were on ferritic welds. Other issues might have arisen in the case of austenitic welds, where there can be additional complications for inspection and qualification.

The application of the direct judgement guidelines was straightforward. There was reasonable consistency between the POD judgements of the individual qualification body members, both at 80% confidence and at 95% confidence.

The use of the relationship between POD and margin of detection was not considered useful for the pilot applications since failure to detect defects was unlikely to be due to low signals. Such failures were instead judged to be more likely associated with aspects such as human factors, surface condition, uncertainties over whether worst case defects had been adequately defined etc. However it was concluded that this relationship could still be useful when the margin of detection is low, as is often the case for austenitic welds.

The trials of the Bayesian approach highlighted a number of issues. These included some which have now been addressed in the Annex 1 guidelines, and some which require further work (or at least should be drawn to the attention of those who intend applying the methodology). The main issues were:

1) It was evident that there were different interpretations among the qualification board members of the meanings of TJ element score and weight, and (in the case of the 2nd pilot) widely differing views on the relative weights of the TJ and trials. It was concluded that careful guidance is required (e.g. in written guidelines, during training, when facilitating the application of this method) to ensure correct and consistent interpretation and application of the method. The Guidelines in Annex 1 address these issues as far as practicable. However it is ultimately a matter of expert judgement what scores / weights are attributed – it is not possible to be prescriptive on this.

2) A TJ can contain two different types of independent information, which must be treated differently.

One type is “stand-alone” information from which POD can be inferred (even if with low confidence). Examples are previous trials results, and results from modelling. If each of these independently indicates a POD of e.g. 90%, these elements support each other so there is increased confidence that POD is 90%.

The other type is “limiting factor” information. Examples are reduction in POD due to coverage, and reduced reliability due to human factors. Reductions in POD due to each of these are cumulative, e.g. if coverage is only 90% and personnel only identify and correctly interpret defect signals 90% of the time, these two factors will be cumulative and will limit POD to a maximum of 81%.

The Guidelines in Annex 1 address this issue.

3) Ultrasonic detection capability depends on a variety of defect features such as orientation, roughness, location etc. It can be difficult to conclude a POD for the overall defect population without information on the relative likelihoods of different classes of defects being present. TJs and practical trials often concentrate on “worst case” defects, i.e. those predicted to be most difficult to detect. Determining POD based solely on these can lead to undue pessimism for the defect population as a whole. It may be appropriate to determine separate PODs for separate populations of defects, but knowledge of the overall POD would still require knowledge of the relative proportions of these populations.

4) Careful consideration is required regarding what is counted as successful detection in practical trials results. If the inspection system is based on application of several scans and defects are detected by more than one scan (e.g. 6 defects each detected by each of 4 different applied probes) a decision is required on whether this represents 6 successful detections or 24.

5) Consideration is required regarding whether / how to take account of the margin of detection for defects during trials. In the current methodology, there is no distinction between only just detecting defects in trials, and detecting them very easily (high signal to report threshold).

6) The methodology does not address accuracy of sizing (a defect could be detected but not reported if undersized) or how to treat false calls.

11.4 Forum for Discussion

There have been very useful discussions on the approaches investigated and the guidelines developed throughout the project. These discussions have involved:

- The three organisations mainly involved in performing the work (Doosan Babcock, JRC, VTT)
- The members of the pilot qualification body
- The sponsors of the project
- ENIQ Steering Committee and TGR

12 RECOMMENDATIONS

1. When quantifying the outcome from inspection qualification it is important to consider carefully the full population of defects to be detected. Technical justifications and practical trials may concentrate on “worst case” defects, i.e. those which are predicted to be most difficult to detect. The POD for this category of defect is likely to be lower than the full population (worst case + non worst case). A POD based solely on consideration of worst case defects may therefore be an unduly pessimistic judgement of overall POD. There may be occasions where it is useful to consider POD separately for different categories of defect.

2. Defect specifications for qualification usually define ranges for certain parameters of the defect such as tilt, skew, roughness, location, etc. It would be useful to know how these parameters are distributed. For example if the specified skew range is between +5° and -5°, it would be useful to know whether skew is equally likely to be any value within this range, or whether likelihood decreases towards extreme values.

   Having such information would help to determine a more accurate POD as explained under point 1 above. Note that this information would also be useful even if qualification outcome is not being quantified, because excessive effort can be involved in trying to detect defects at worst case combinations of skew, tilt, roughness etc when in fact the likelihood of such defects existing may be vanishingly small.

   It is therefore recommended that work is undertaken to study how such defect parameters are likely to be distributed in reality.

3. Application of the direct judgement process to determine POD as a step curve is relatively straightforward. In principle the simplification of the POD curve in this manner is no different from the common practice of qualifying against a single defect size with no explicit requirements for detection of defects below that size.

   However it can still be useful to consider the relation between risk reduction and details of the POD curve, since even in the case of a step curve this can help achieve an appropriate balance between risk reduction and the difficulty of meeting specific qualification targets (defect qualification size and POD).

4. Further work to establish a more rigorous relation between margin of detection and POD is recommended since this would be a useful tool to support direct judgement in the case where detection margin is low e.g. austenitic welds (stainless steel and Inconel) or complex geometries – see Section 6 and Ref (5).

5. Linking POD, risk reduction and inspection interval requires knowledge of defect initiation and growth. Sections 8 and 10 of this report illustrate this process using relatively simple models and assumptions. For formal application, more detailed models are likely to be required including consideration of repair criteria (a simplifying assumption was that all detected defects were repaired).

6. The guidelines in Annex 1 provide a good starting point for those considering application of the Bayesian approach. However further work is recommended to make the application more robust in practice. In particular the issues identified in Section 11.3 should be addressed and this may require further pilot applications.

7. Consideration should be given to how to construct a TJ to facilitate quantification. This could involve for example organising the TJ so that the “stand-alone elements” and “limiting factors” (see Section 11.3 and Annex 1) are explicitly identified and discussed.

8. It is recommended that opportunities are sought to trial the quantification approaches described in parallel with formal qualification exercises on real plant. Even if the quantification data generated are not used for formal purposes, the lessons learnt from such trials would be extremely useful.

9. ENIQ could play an important role in taking forward some of these recommendations. In particular further work is recommended on the Bayesian Guidelines in Annex 1, with a view to them potentially becoming a formal ENIQ document. Such work could be performed jointly by an ad-hoc group of members from ENIQ TGR and TGQ.

13 REFERENCES

Ref. 1 European methodology for qualification of non-destructive testing: third issue, ENIQ report. 31, EUR 22906 EN, 2007.

Ref. 2 European Framework Document for Risk-informed In-service Inspection, ENIQ report 23, EUR 21581 EN, 2005.

Ref. 3 Group Sponsored Project 238 Proposal "Link Between Risk-Informed In-service Inspection and Inspection Qualification". Doosan Babcock May 2006

Ref. 4 Gandossi, L. & Simola, L., Technical Report "Sensitivity of Risk Reduction to Probability of Detection Curves (POD) Level and Detail" ERU Report 22675 EN, May 2007

Ref. 5 Shepherd B, Goujon S, Whittle J "Link Between RI-ISI and Inspection Qualification: Relationship between Defect Detection Rate and Margin of Detection". SKI report 2007:04

Ref. 6 Gandossi, L. and Simola, K., Framework for the quantitative modelling of the European methodology for qualification of non-destructive testing, International Journal of Pressure Vessels and piping 82 (2005) 814-824.eport

Ref. 7 Gandossi, L. & Simola, L., Technical Report "A Bayesian framework for the quantitative modelling of the ENIQ methodology for qualification of non-destructive testing" EUR Report 21902 EN, May 2007.

Ref. 8 Fleming, K.N. Markov models for evaluating risk-informed in-service inspection strategies for nuclear power plant piping systems. Reliability Engineering & System Safety, Volume 83, Issue 1, January 2004, Pages 27-45.

Ref. 9 Simola, K, Mengolini, A, Bolado-Lavin, R, "Formal Expert Judgement – an Overview" EUR Report 21772 EN, July 2005.

ANNEX 1: GUIDELINES FOR APPLICATION OF BAYESIAN METHODOLOGY

1) Organisation as expert elicitation process.

1.a) It is recommended that the quantification of a Technical Justification is organised as a structured expert elicitation exercise.

The process of formal expert judgement usually consists of the following steps:

1) Definition of the issues about which the expert judgements should be made.
2) Training of the experts and definition of variables to be elicited.
3) Individual work of experts.
4) Elicitation (drawing out the opinions of the experts)
5) Analysis and aggregation of results and, in case of disagreement, attempt to resolve differences.
6) Documentation of results, including expert reasoning in support of their judgement.

More information and additional references can for instance be found in [Ref 9]

1.b) The panel should have at least 4 members

In order to ensure a sufficiently broad representation of expert opinion, it is recommended that the panel has at least 4 members. One of them will be the facilitator who should be knowledgeable in the field of subjective probability, and should have a thorough knowledge of the methodology, of the meaning of all the relevant concepts and of all its potential shortcomings. The facilitator should be satisfied that the other panel members have received sufficient training in the methodology, and is responsible for the process of elicitation, aggregation and final reporting of the case. It is preferable that the facilitator does not contribute an expert judgement in order to ensure objectivity during the elicitation process

One panel member should act as a chairperson. The main role of the chairperson is to help the panel achieving a consensus.

1.c) The whole process should be thoroughly documented.

The documentation should include, as a minimum, a description of the panel participants (name and summary of the relevant expertise), minutes of all the relevant discussions, the quantities elicited, and a thorough description of the method used to combine the judgments of the various experts.

2) Training

2.a) The following topics should be covered:

- Short introduction to expert judgement: what it is, how it works, and what are the common biases encountered when experts are giving judgements under uncertainty.
- General concepts of probability.
- Bayesian methodology for the quantification of ENIQ methodology.
- A worked example of the Bayesian methodology
- An exercise on the elicitation of uncertain quantities.

2.b) It is essential that the technical experts are thoroughly briefed regarding what kind of judgements they will required to give and how the model works.

In particular, concepts such as "independent element", "weight" and "score" can be prone to misunderstandings. We therefore recommend that an adequate amount of time during the training is spent to make sure that there is a correct (and common) understanding among the technical experts.

2c) Extent of training

Technical experts should have had at least two full days training to ensure the above are covered sufficiently.

3) Recommended method for the quantification of an ENIQ TJ

3.a) “Approach 1” described in [Ref. 7] is recommended. The main steps are as follows.

1. The approach is based on quantifying the outcome from inspection qualification using Bayesian statistics. The purpose is to produce a quantitative estimate of the probability of detection based on the combined information of the TJ and practical trials.

2. The fundamental idea is to treat the TJ as a set of “equivalent” practical trials. Thus, all the relevant evidence on inspection capability contained in the TJ is seen as actually equivalent to performing a number of trials.

3. The probability of detection *p* (for a specified population of defects, see "4 - Defect population & splitting of the evaluation") is described using a Beta-Binomial model. In other words, finding a defect is modelled as a Bernoulli trial, with success probability *p*. In turn, the parameter *p* is modelled as a random variable and the uncertainty related to it is expressed with a Beta probability distribution.

   $$p \sim Beta\ (\alpha,\ \beta)$$

   α and β are the parameters of the Beta distribution, describing the current knowledge about *p*.

4. If an experiment is carried out, consisting of a set of *N* trials (i.e. the NDE system at hand is applied to a set of N flaws), N_s of such flaws will be detected (successes) and $N_f=N-N_s$ will be missed (failures). A new (posterior) distribution can be obtained by simply adding N_s to α and N_f to β.

   $$\alpha_{post}=\alpha_{prior}+N_s$$

   $$\beta_{post}=\beta_{prior}+N_f=\beta_{prior}+N-N_s$$

5. The TJ is interpreted in terms of an equivalent set of practical trials. In Approach 1, the TJ is quantified using two numbers (see Figure next page):

- an equivalent total number of trials, N_{TJ}, and
- an equivalent number of successes, $N_{TJ,s}$.

6. These numbers are elicited from the experts in a documented and transparent manner, and are then used in combination with the number of practical trials, N_{trials}, (and associated number of successes, $N_{trials,s}$) to determine (or prove, according to the specific application) the achievement of the qualification capability (or objectives).

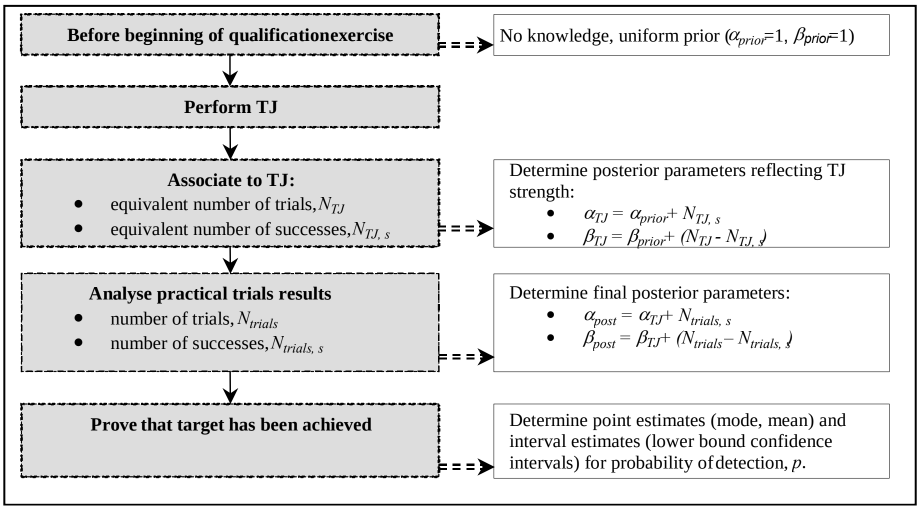

Before beginning of qualificationexercise
No knowledge, uniform prior (α_{prior}=1, β_{prior}=1)
Perform TJ
Associate to TJ:
• equivalent number of trials, N_{TJ}
• equivalent number of successes, $N_{TJ,\ s}$
Determine posterior parameters reflecting TJ strength:
• $\alpha_{TJ} = \alpha_{prior} + N_{TJ,\ s}$
• $\beta_{TJ} = \beta_{prior} + (N_{TJ} - N_{TJ,\ s})$
Analyse practical trials results
• number of trials, N_{trials}
• number of successes, $N_{trials,\ s}$
Determine final posterior parameters:
• $\alpha_{post} = \alpha_{TJ} + N_{trials,\ s}$
• $\beta_{post} = \beta_{TJ} + (N_{trials} - N_{trials,\ s})$
Prove that target has been achieved
Determine point estimates (mode, mean) and interval estimates (lower bound confidence intervals) for probability of detection, p.

Figure A1 Principle for combining evidence from TJ and practical trials to prove that reliability target is achieved.

7. Approach 1 is based on quantifying the technical justification in terms of score and weight.

   If some evidence is missing, this should imply less weight for the TJ. This means that the equivalent TJ sample size, N_{TJ}, should be smaller than in the case of stronger evidence.

   If evidence is present showing that some defects could be missed, this should imply a lower expected score of the TJ, i.e. the ratio of $N_{TJ,s}$ over N_{TJ} should be smaller than in the case where the evidence is more supportive regarding detection capability.

8. The problem is thus reduced to the determination of the two numbers N_{TJ} and $N_{TJ,s}$ and how to weight the evidence in the TJ with respect to the trials.

   The user must be particularly aware of the fact that these two numbers are ultimately a matter of expert judgement.

   If the expert(s) involved in the process feel confident enough, they can provide such numbers directly. However, it is recommended to break down the TJ into elements, as suggested in [Ref. 6] and [Ref. 7].

9. If the TJ is broken down into elements, the next step is to define weighting and scoring principles for the quantification. Weighting is associated with the determination of N_{TJ}, i.e. the relative weight that the technical justification

has when compared with the practical trials. Scoring is associated with the determination of $N_{TJ,s}$, i.e. the judgement of the extent to which the TJ supports 100% detection

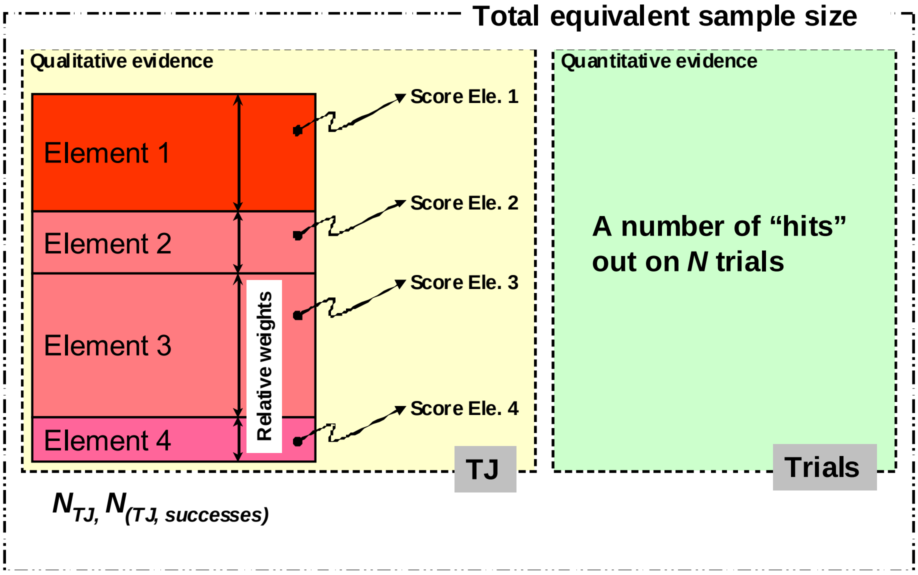


Figure A2 Weighting and scoring of TJ elements

10. The final posterior parameters are obtained as

$$\alpha_{post}=1 + N_{TJ,s} + N_{trials,s}$$

$$\beta_{post}=1 + (N_{TJ} - N_{TJ,s}) + (N_{trials} - N_{trials,s})$$

And the knowledge about p is fully expressed by the posterior distribution:

$$p \sim Beta(\alpha_{post}, \beta_{post})$$

This function can then be used to extract the required information (mean, mode, confidence intervals, etc.)

3.b) The user should be particularly aware of the complexities related to identifying the elements in which the TJ is to be broken down, see "5 - Selection of elements" below.

4) Defect population & splitting of the evaluation

4.a) The overall defect population for which POD is required will generally be all defects which meet the qualification specification. Thus the population typically covers all defects exceeding specified dimensions and which lie within specified ranges for tilt, skew, location, morphology etc. It should be borne in mind however that certain categories of defect (in terms of how these parameters combine) are likely to be easier to detect than others. For example technical justifications and practical trials often concentrate on worst case defects, i.e. those predicted to be most difficult to detect. The POD for this category of defect is likely to be lower (by definition) than for the whole population (worst case + non worst case). A POD based solely on consideration of worst case defects may therefore be an unduly pessimistic judgement of overall POD.

4.b) There may therefore be occasions when it is useful to split the overall population into different categories of defect and to determine POD separately for each. Note however that combining the results to derive an overall POD will require prediction of the relative proportions of defects within these different categories. This in turn will require knowledge or assumptions regarding how parameters such as tilt, skew etc are distributed within the specified tolerance bands. It is likely to be more appropriate for the combination of PODs for different categories of defect to be made by structural reliability specialists rather inspection qualification personnel.

5) Break down of TJ and selection of elements

5.a) In principle the judgement of the TJ in terms of an equivalent number of successful trials and successes could be made in a single step. However it is recommended that this judgement is made in a structured and transparent manner by breaking down the TJ into separate independent sources of evidence. These independent sources of evidence can fall into one of two categories: stand-alone elements (5.b) and limiting factors (5.c).

5.b) Stand-alone elements.

One category is “stand-alone” information from which POD can be inferred (even if with low confidence). Examples are previous trials results, and results from modelling. If each of these independently indicates a POD of e.g. 90%, these elements support each other so there is increased confidence that POD is 90%.

Each of the stand-alone elements within the TJ should be given a score corresponding to the extent to which it points towards 100% detection, and a weighting which reflects the relative value of the evidence it provides. For example if evidence from modelling suggests that 9 out of 10 defects will be detected the score would be 0.9 and if this modelling evidence is judged to contribute (in terms of value of the evidence) 40% of the total evidence within the TJ, its weight would be 0.4. The weighted score for this element would then be 0.9 x 0.4 = 0.36.

The weighted scores for each of these stand-alone elements are then added (Note that total weight must = 1). This then represents a score for the TJ (maximum value = 1) before taking into account the limiting factors discussed next.

5.c) Limiting factors.

Limiting factors are elements that can, alone, set an upper limit on the probability of detection. Such elements must be treated separately.

A limiting factor could be coverage. For instance, it might be deduced from the information contained in the TJ that coverage is only 75%. This will automatically reduce the maximum attainable probability of detection to 0.75, as all defects located in the 25% of the volume not covered by the NDE system will be missed (assuming uniform distribution).

Another limiting factor might be equipment. For instance, it might work satisfactorily only 95% of the time. Again, the maximum probability of detection could not be higher that this value.

Limiting factors are cumulative, so that if coverage alone is responsible for missing 25% of defects and equipment alone is responsible for missing 5%, then these two factors together are responsible for missing 29% of defects ($1-0.75 \cdot 0.95$) 100% compared to the situation if those two factors had been perfect.

Let p_{LF} be defined as the cumulative probability of missing a defect due to limiting factors. In the example above $p_{LF} = 29\%$.

5.d) Having broken the TJ down into various stand-alone elements and limiting factors, they are combined as follows:

1. Obtain the posterior distribution as from step 10 of point 3.a, taking into account stand-alone elements only. This is a Beta distribution function defined in the interval [0, 1], i.e. it covers the POD probability distribution for in the range 0% to 100%;

2. Determine the cumulative probability, p_{LF}, of missing a defect due to limiting factors (last paragraph of point 5.c). This expresses the fact that the limiting factors impose an upper limit equal to 1-p_{LF} on the POD;

3. In effect the Beta distribution applies to the reduced population of defects which are within the area covered, and where equipment works, etc. If for example, p_{LF}=29%, the combined limiting factor, 1-p_{LF}, is 71% and then the initially derived Beta distribution must be scaled so that it only covers POD in the range 0% to 71%. Beyond 71% it is zero.

4. Figure A3 provides an illustration of the approach. The final distribution has the same shape as the one obtained quantifying the stand alone elements only, but "squeezed" between 0 and 1-p_{LF}. Consequently, it is also scaled in the y-axis direction by a factor $1/p_{LF}$, so that the total area under it remains equal to 1. All the relevant quantities, such as mean, mode and percentiles are simply scaled by a factor 1-p_{LF}.

5.e) Importance of independence

It is important to ensure that the various components of evidence are independent, whether of the stand-alone or limiting factor type. For example the POD from the element “previous experience” might be only 80% specifically because of the limiting factor “unreliability of personnel”. Similarly results from modelling might already take into account restrictions in coverage. Ensuring that the various components of evidence are independent will avoid double accounting.

6) Weight of the TJ

6.a) The "weight" of the evidence contained in a TJ as expressed by means of the equivalent TJ sample size, N_{TJ}.

Providing a judgement on such number can be a difficult task for the experts.

The experts can be aided in this task by visualising how many practical trials with representative defects (of the types for which detection is to be quantified) they would be ready to accept if they were to "give up" the information supplied by the TJ in support of inspection capability in order to maintain the same degree of understanding in the NDE system.

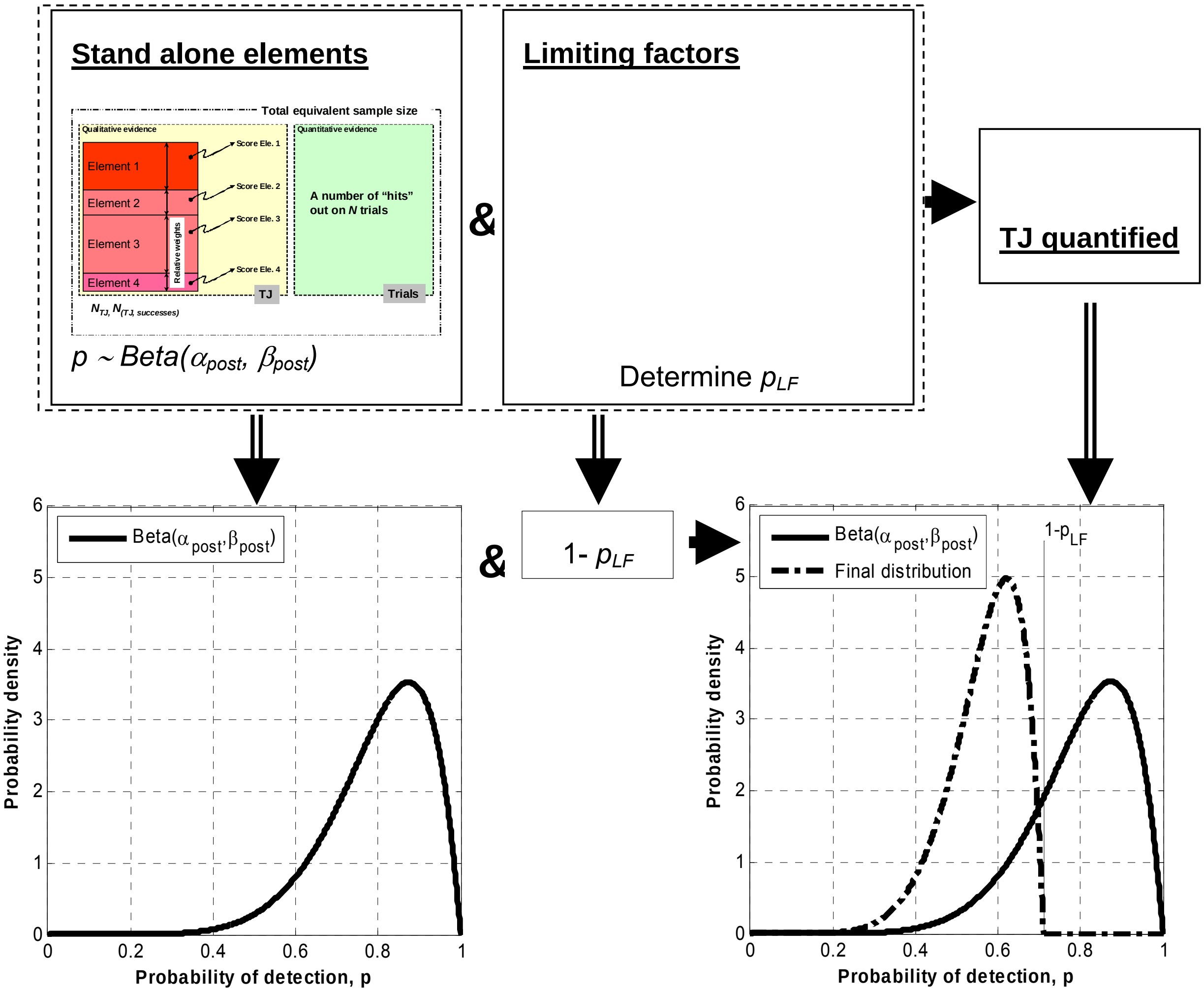


Figure A3 Model for the combination of stand alone elements and limiting factors

7) How to define one practical trial.

7.a) The definition of a practical trial can be problematic.

The model simply requires that each practical trial be assumed as a Bernoulli trial, i.e. as an independent trial whose output is either a "hit" (success) or a "miss" (failure). Generally components are inspected with multiple probes and this is simulated during practical trials. For instance, the tespiece(s) might contain 6 defects. Each defect might then be inspected with 6 different probes, thus giving 36 defect-beam interactions. If each interaction produces a reportable signal, the question then arises whether the practical trials can be treated as 6 out of 6 hits, or 36 out of 36.

It could happen that only 1 of the 6 beams produces a reportable signal when scanning a particular defect, and still the defect would be reported as detected. In this case, assuming 5 failures out of 6 trials would not be representative of the real NDE system capability. The correct assumption in this case would be to treat the case as 1 success out of 1 trial (but see also point 8.c below).

It is thus recommended that the user satisfies him/herself that a single (and independent) practical trial is defined in a way that is truly representative of the situation that will happen when the NDE system is applied in practice.

8) Weighting for special cases

8.a) Worst-case defects in practical trials.

When test blocks are designed, the aim is often to manufacture defects that represent the most difficult defects to be found. The user may decide to take some credit for this situation.

One would expect that actually the probability of detection (e.g. of defects exceeding a certain size) is higher in the case where the defect population includes not only the most difficult defects (due to their tilt, or skew etc.), but also the whole range of possible defects.

Thus, it could be claimed that finding 10 out of 10 test piece “worst case” defects actually results in higher confidence than what a statistical analysis (with the assumption of identical defect populations) indicates.

In [Ref. 7, Section 6.3], a mathematical framework was suggested to deal in a rigorous way with this situation, partitioning the population of defects into two groups each characterised by a different probability of defects. To draw practical conclusions, the user would then be required to make further assumptions, for instance about the relative percentage of worst-case defects as opposed to the rest of the defect population.

An alternative approach would be to simply give a higher weight to worst-case practical trials, for instance counting each of them as being equivalent to performing two trials. This decision must be taken (and justified) by the user, with the full awareness that it remains based (like the rest of this approach) in expert judgement.

8.b) Margin of detection.

A similar line of reasoning as above can be made regarding margins of detection, if the user feels justified taking some credit for the fact that defects in test piece trials are detected with very high margins above the reporting threshold.

8.c) Number of beams.

A similar line of reasoning as above can be made regarding number of beams, if the user feels justified taking some credit for the fact that defects in test piece trials are simultaneously detected by multiple beams.

9) Worked example

In this example, the qualification body is asked to quantify detection capability of an NDE system which is being qualified. A TJ is considered, in which a body of evidence supporting detection capability has been assembled. Further, 10 practical trials have been carried out. The NDE system has correctly found all 10 flaws.

Step 1 – Identification of stand-alone elements and limiting factors.

Evidence is present in the TJ from which it is concluded that coverage is limited to 95% of the inspection volume. Further it is judged that equipment may work satisfactorily 99% of the time and that human factors cannot be higher than 98%. These three elements represent limiting factors. Three stand-alone elements are then identified within the TJ: (1) theoretical modelling, (2) experimental evidence and (3) parametric studies.

Step 2 – Determination of cumulative probability of missing a defect due to limiting factors, $\mathbf{p_{LF}}$.

$$p_{LF} = (1\text{-}0.95\cdot 0.99\cdot 0.98) = 0.078$$

In other words, the maximum probability of detection will not be higher than $1\text{-}p_{LF}$ = 92.2%

Step 3 – Decision regarding the relative weights of the TJ stand alone elements.

The first decision the qualification body is asked to make concerns the relative weights of the TJ stand alone elements. In the Technical Justification at hand, three main stand alone elements have been identified. After examination, the qualification body decides that Element 1 contributes towards 30% of the total evidence contained in the TJ, Element 2 is judged to carry 50% of the evidence, and Element 3 20%. These values are summarised in the first column of Table A1, as fractions of 1. The sum of these contributions must necessarily be 1.

Step 4 – Decision regarding the score of the TJ stand alone elements.
The next step consists of scoring the TJ stand alone elements. It is important to reiterate the notion that this score must reflect how well the evidence contained in the TJ element supports the detectability of the prescribed defects.

In the example, Elements 1 and 3 are judged to fully support the detectability of all defects in the specified population, and are thus both assigned a score of 100%. Considerations in Element 2 indicate that some limiting defects (such as worst case combinations of size, tilt and skew) could very occasionally be missed. Element 2 is thus scored with a 95%, expressing an intuitive notion that roughly 1 defect in 20 could be missed (purely according to the evidence contained in this Element).

These values are reported in the second column of Table A1, again as fraction of 1.

Step 5 – Calculation of TJ total weighted score.
The score of the TJ as a whole is easily obtained. For each element, the score and weight are multiplied and an element weighted score is obtained (third column of Table 6). The elements' weighted scores are finally added together to determine the TJ total weighted score, w_{TJ}.

In the example (Table A1), according to the weight and evaluation given to the elements, the total TJ score is estimated to be 0.975.

Table A1 Hypothetical data used in the example.

	Relative weight	Score	Weighted score
Element 1	0.3	1	0.3
Element 2	0.5	0.95	0.475
Element 3	0.2	1	.2
	$\Sigma = 1$		**Σ = 0.975= w_{TJ}**

Step 6 – Decision of the TJ equivalent sample size.
The next decision concerns the TJ equivalent sample size, N_{TJ}. In other words, it must be decided how much the technical justification is weighted as a whole. A possible way forward is to weight the TJ directly against the number of practical trials.

10 practical trials have been performed. The qualification body proceeds by trying to determine how many equivalent trials the TJ is worth in comparison with this. In other words, the qualification body tries to decide how many additional practical trials it would like to add to the original 10 in order to gain the same confidence that is given by the evidence contained in the TJ.

The qualification body is specifically asked to make an expert judgement. It thus needs to turn to the TJ and examine it with a critical eye. In the example, the qualification body analyses the TJ and realises that a lot of resources have been invested in assembling several pieces of evidence, all of which seem to contribute in building a very high confidence that all the required defects will be detected.

The qualification body eventually decides that just about 20 additional practical trials would yield the same confidence if it was asked to give up the evidence contained in the TJ. Thus, it is decide that:

N_{TJ} = 20

Step 7 – Calculation of TJ posterior parameters (TJ updating)
The TJ total score is used to determine the equivalent number of successes, N_{TJs}, in the following way:

N_{TJs}= $w_{TJ} \cdot N_{TJ}$

The Bayesian updating process is started with a uniform prior, thus we have:

α_{prior} = 1 β_{prior} = 1

The parameters of the Beta posterior distribution (after updating with the evidence provided by the TJ) are thus obtained as follows:

α_{TJ} = 1 + N_{TJs} β_{TJ} = 1 + N_{TJf} = 1 + N_{TJ} - N_{TJs}

The expected value and mode of the posterior distribution are (see [Ref. 7]):

$$E(p) = \frac{\alpha}{\alpha + \beta} = \frac{1 + N_{TJ,s}}{2 + N_{TJ}}$$

$$Mode(p) = \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{N_{TJ,s}}{N_{TJ}} = (w_{TJ})$$

Thus, the mode of the posterior distribution (i.e. the location where the distribution function attains its maximum) is equal to the TJ total score. This is appealing from an intuitive point of view. In the example at hand:

N_{TJs}= $w_{TJ} \cdot N_{TJs}$=0.975×20=19.5

α_{TJ} = 20.5 β_{TJ} = 1.5

E(p)=0.932
Mode(p)=0.975

The probability density function of the TJ posterior is plotted in Figure A4 with a black dashed line.

Step 8 – Updating with evidence from practical trials
The evidence obtained from practical (open) trials can now be taken into account. Since:

N_{trials} = 10 $N_{trials,s}$ = 10

a second posterior is thus easily obtained.

$\alpha_{trials} = \alpha_{TJ} + N_{trials,s}$ $\beta_{trials} = \beta_{TJ} + N_{trials,f}$

In the example at hand:

$\alpha_{trials} = 30.5$ $\beta_{trials} = 1.5$

The expected value and mode of the second posterior are:

E(p) = 0.953
Mode(p) = 0.983

The probability density function of the second posterior is plotted in Figure A4 with a black dash-dot line.

Step 9 – (optional) Updating with evidence from blind trials
If further evidence, obtained for instance from blind trials, was available, a third posterior could be obtained.

Let us for instance assume that the NDE system in the example at hand has been applied to a set of 15 blind trials. We suppose that a single defect was missed. Then:

$N_{blind\ trials} = 15$ $N_{blind\ trials,s} = 14$

A third posterior is again obtained:

$\alpha_{blind\ trials} = \alpha_{trials} + N_{blind\ trials,s}$ $\beta_{trials} = \beta_{trials} + N_{blind\ trials,f}$

Thus

$\alpha_{blind\ trials} = 44.5$ $\beta_{trials} = 2.5$

Expected value and mode of the third posterior are:

E(p) = 0.947
Mode(p) = 0.967

The probability density function of this posterior is plotted in Figure A4 with a red dash-dot line.

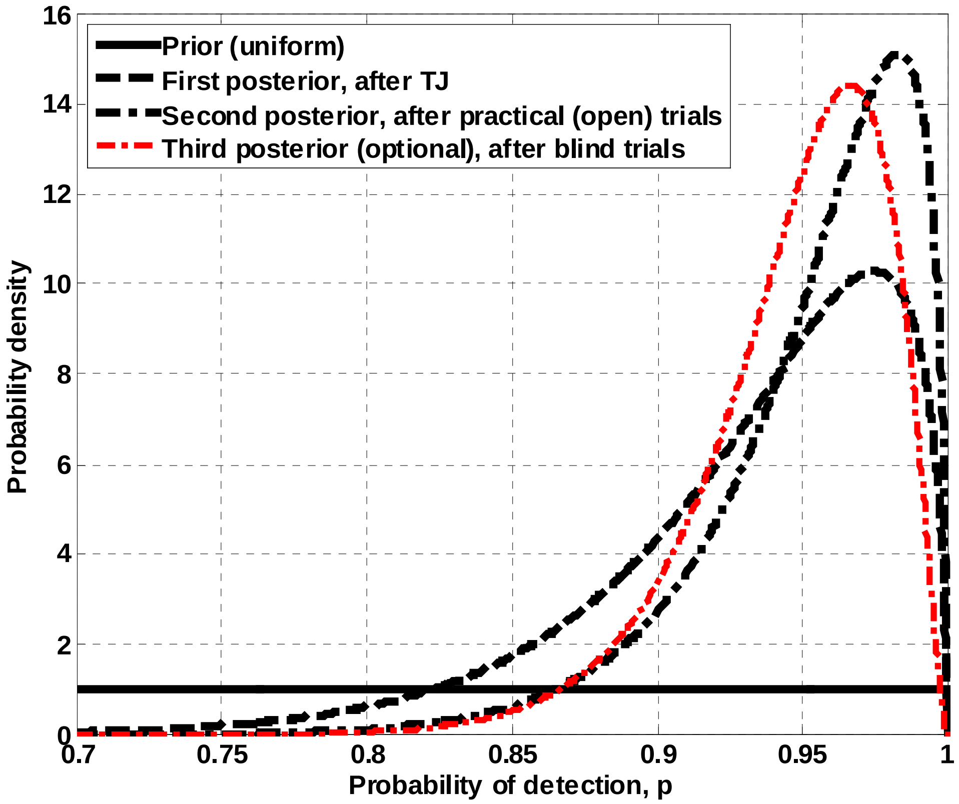


Figure A4 Posterior probability densities for worked example

Step 10 – Determination of final probability distribution (stand alone elements and limiting factors)

The Beta distribution determined at Step 8 (or 9) is then applied to the reduced population, taking into account the reduced area covered, the equipment not always working, and the human factors. We had derived:

$p_{LF} = 0.078$

And thus the final distribution is obtained as described at Point 5.d. Such distribution has the same shape as the one obtained quantifying the stand alone elements only (for instance, at Step 8), but "squeezed" between 0 and $1\text{-}p_{LF}$. The function is scaled in the y-axis direction by a factor $1/p_{LF}$, so that the total area under it remains equal to 1, see Figure A5.

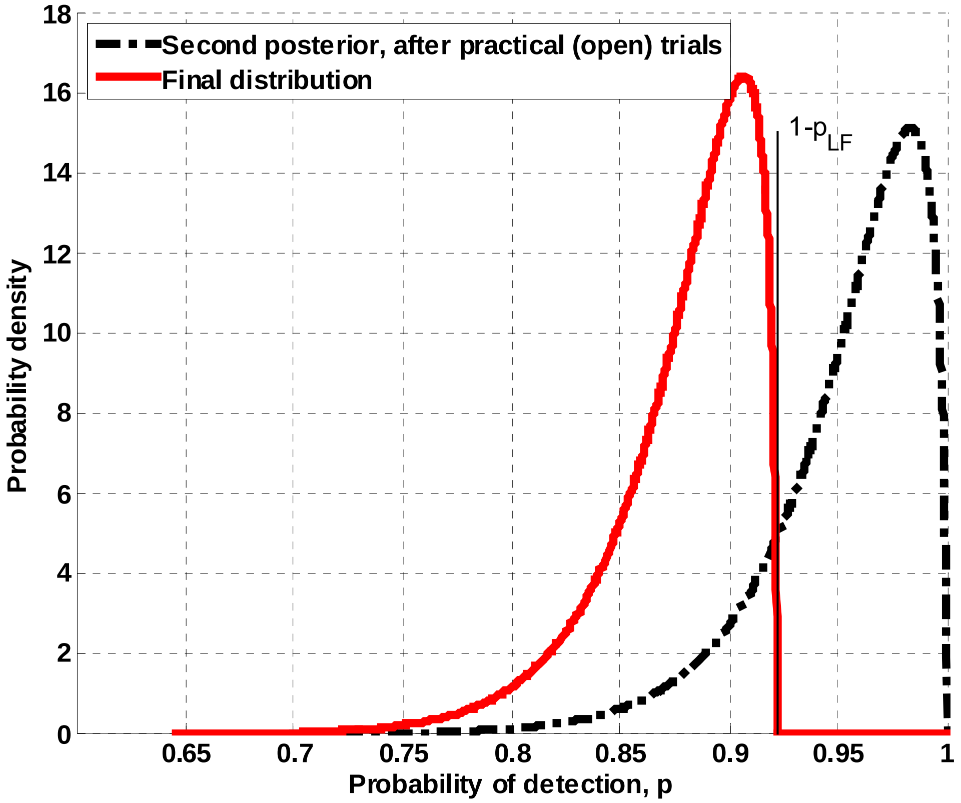


Figure A5 Final probability densities for worked example

The mode of the final distribution is:

$$Mode(p) = (1- p_{LF})\ \ 0.983 = 0.906$$

Step 11 – Reporting

As discussed in [Ref. 6], a convenient way to offer a complete summary of how the information available is described by the posteriors obtained in the updating process is by means of 1-F curves, with F the cumulative Beta distribution function. In these curves, plotted in Figure A6, the abscissa *x* represents the lower bound probability of detection, *p*, and the ordinate y represents the associated confidence level, δ. Therefore, the curves of Figure A6 allow obtaining directly, for any given required confidence level, the correspondent probability of detection.

The final distribution is plotted with a solid red dot line. At a confidence level of 90%, the lower bound probability of detection is found to be just approximately 0.87.

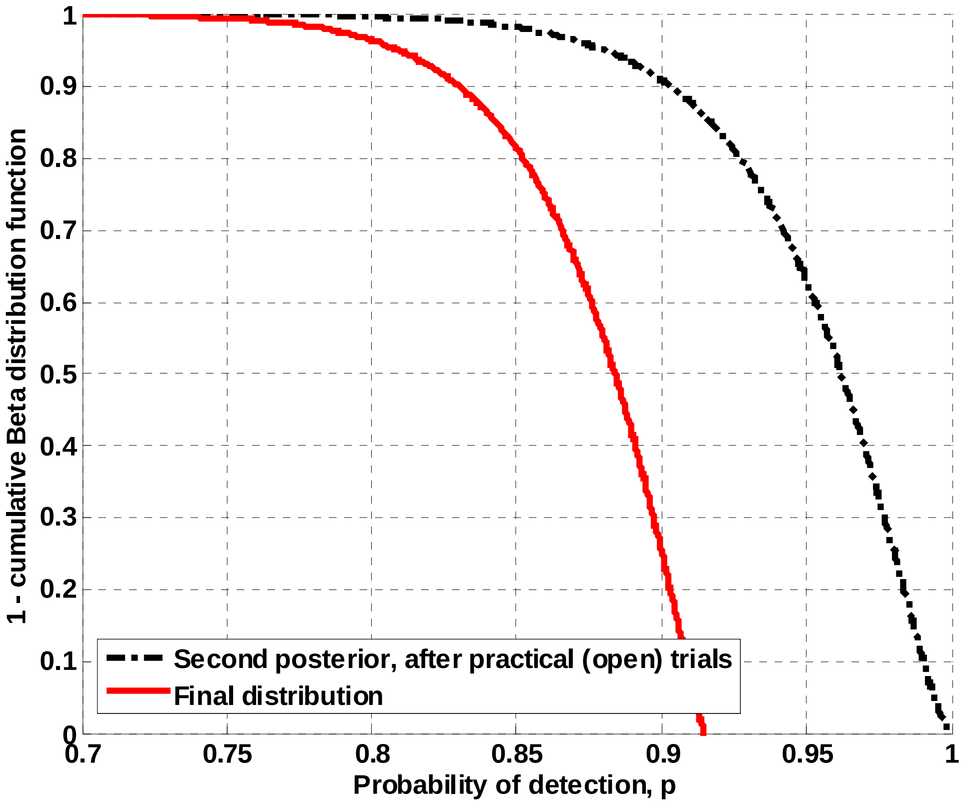


Figure A6 1-F curves for worked example

European Commission

EUR 23928 EN — DG JRC — Institute for Energy

ENIQ TECHNICAL REPORT: LINK BETWEEN RISK-INFORMED IN-SERVICE INSPECTION AND INSPECTION QUALIFICATION

Authors

Barrie Shepherd	Doosan Babcock
Luca GANDOSSI	DG-JRC-IE
Kaisa SIMOLA	VTT Technical Research Centre of Finland

Luxembourg: Office for Official Publications of the European Communities
2007 – 22 pp. — 21 x 29.7 cm
EUR — Scientific and Technical Research Series; ISSN 1018-5593

Abstract

There is a growing need for a quantitative measure of inspection effectiveness as an input to quantitative risk-informed in-service inspection (RI-ISI). A Probability of Detection (POD) curve could provide a suitable metric. However there can be significant problems associated with generating realistic POD curves by practical trials. The ENIQ inspection qualification methodology can provide high assurance that an inspection system will achieve its objectives, but is not designed to provide a quantitative measure of the type that can be used in RI-ISI analysis.

A project, led by Doosan Babcock, was therefore set up with main objectives to investigate approaches to quantifying the confidence associated with inspection qualification and to produce guidelines on how to relate inspection qualification results, risk reduction and inspection interval.

This report discusses how a simplified POD curve, such as a step curve, could be used as the target for inspection qualification, or as an output from it. Work to investigate the sensitivity of relative risk reduction to the details of the POD curve is described from which it is concluded that use of a simplified POD curve could be justified.

Two methods for quantifying the outcome from inspection qualification are described. The first method is a relatively simple process based on direct expert judgement. The second method is based on a more rigorous structured process employing Bayesian statistics, in which the subjective degree of belief in inspection capability derived from a Technical Justification (TJ) is expressed in probabilistic terms, and combined with data from practical trials results. Two pilot studies are described which involved a qualification body applying the quantification methods in practice. Recommendations for further work to make the approaches developed more robust are provided.

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Publications Office
Publications.eu.int