Reliability Investigations of Radiographic Testing
Using aRTist as a Simulation Tool

Christian GOLLWITZER $^1$, Carsten BELLON $^1$, Andreas DERESCH $^1$, Gerd-Rüdiger JAENISCH $^1$, Hans-Uwe BARON $^2$, Uwe EWERT $^1$

$^1$ BAM Federal Institute for Materials Research and Testing; Berlin, Germany
Christian.Gollwitzer@bam.de
$^2$ MTU Aero Engines GmbH; Munich, Germany; Hans-Uwe.Baron@mtu.de

Abstract
Within the framework of the European project PICASSO, the radiographic simulator aRTist (analytical Radiographic Testing inspection simulation tool) developed by BAM has been extended for reliability assessment of film and digital radiography. A simulation supported probability of detection (POD) methodology has been developed and the validity of the approach has been studied using an application from the aeronautics industry. An experimental POD has been determined with the help of a specialized software tool, developed to aid with the collection of large series of POD data. The resulting POD is compared to simulations of the same setting using aRTist and the newly available module for simulation supported POD. A quantitative agreement within a few percent is achieved between the experimental and the simulation supported POD.

Keywords: X-ray Imaging, Probability of Detection, Simulation

1 Introduction
Reliability investigations of NDT methods traditionally involve NDT trials, which are series of tests where test pieces with known flaws are examined using the method in question [1]. The performance of the method is evaluated afterwards by comparing the findings of the test to the known truth. The desired final output of this procedure is the probability of detection (POD) of the flaws. The most commonly used procedure to determine this POD($\alpha$) as a function of the flaw size $\alpha$ by experiments has been established by Berens [2] more than half a century ago. Today this concept is widely accepted by many industrial sectors, e.g. railway, aircraft, power plants, and the petrochemical industry. While the originally proposed method is a very direct way to deduce the capabilities of an NDT procedure from the experimental data, it has the drawback that a large number of test pieces is required to get a good estimate of the POD. These test pieces are expensive to manufacture and additionally must be characterised precisely to know the truth about the defects. With the advent of realistic NDT simulation software it has been proposed to perform simulations instead of real experiments and a lot of research has been carried out on how to match these simulations with real NDT data. Specifically, the US MAPOD working group [3] and the European project PICASSO [4] have made many efforts towards the development of such concepts. PICASSO also aims at providing usable software to perform simulation supported POD, and data from specific test cases to check the validity of the software.

Within this context the radiographic simulator aRTist [5] (analytical Radiographic Testing inspection simulation tool) developed by BAM has been extended by the module SimuPOD, which facilitates reliability assessment of film and digital radiography [6], but has not yet
been extensively checked against real data. This paper describes the first comparison of aRTist’s model to an RT test case from PICASSO. The paper is organized as follows: first, the POD facilities of aRTist are briefly reviewed. Then, the RT test case and the experimental data generation are described. Finally, the POD curves obtained by simulation and experiment are compared and conclusions are drawn.

2 aRTist and SimuPOD

aRTist is a software package for the simulation of film and digital radiography [5]. The user can specify the radiographic setup, including position and orientation of the source and detector, in an interactive fashion by direct manipulation in a 3D view. Simple objects can be constructed directly in aRTist from primitives like cuboids, spheres, and cylinders using Boolean operations and complex objects can be read from triangulated CAD exchange files. All user-specified parameters are related in a straight-forward way to the real setup; for example, the material of objects is chosen from a predefined list of often used materials and new material definitions can be added by entering the respective chemical composition. Similarly, the source spectrum and brightness can be computed for an X-ray tube by entering the voltage, current, and filtering applied. The detector can be chosen from a predefined list of radiographic films, imaging plates, and digital detectors. The employed model respects the spatial resolution and spectral sensitivity of a given detector. The interaction of the photons with matter is modelled using polychromatic attenuation only. Some simple approximations to estimate the scatter radiation are also included, but have not been used in the simulations described in this document. aRTist is very fast – during the interactive manipulation a live preview is generated which simplifies the construction of the virtual setting.

To facilitate an easy workflow for the computation of POD data, aRTist has been enhanced with the module SimuPOD [7], which adheres to the POD($a$) model. The user must first set up a radiographic scene containing defects. Then, the defects are marked and a variation schedule is defined. This comprises the geometric transformation of the flaw size $a$ in the range of interest plus additional random fluctuations, and the desired number of images. This series of simulations is run in batch and after it is finished an automatic image processing evaluates the visibility of the defects in the resulting images. An interactive analysis module performs the $\hat{a}$-vs-$a$ calculation following Berens [2] and the resulting POD($a$) curve is displayed and the critical thresholds $a_{90}$ and $a_{90/95}$ are computed. Any intermediate results – namely the images, the $\hat{a}$-vs-$a$ data, and the POD($a$) curve – can be exported in standard formats to enable further processing by other means.

Up to this point the POD calculated using this method is purely computational – no experimental data has been used in the construction. To solve this issue, a realistic test case has been defined by PICASSO, which is used to verify and adjust the simulation results. The experiments with this test case are described in the next section.
Experimental POD trials

The validation case deals with the detection of pores in turbine disks made of forged Ti6246 alloy. The expected size of the defects is in the range of 50 µm to 400 µm and the wall thickness ranges from 3 mm to 10 mm. Since a real turbine disk is quite expensive and imposes additional constraints on the X-ray geometry, the experiments have been carried out using a test piece dummy consisting of solid wedges and sheet metal with drilled holes.

3.1 Image acquisition

The geometry is depicted in figure 1. There are two wedges with thickness ranges from 3–7 mm and 7–10 mm on a length of 160 mm and a width of 60 mm. The two metal sheets with thicknesses of 50 µm and 100 µm are made of pure Titanium and contain 5 double rows of drilled holes each with different nominal diameters of 50, 100, 200, 300, and 400 µm and 25, 50, 100, 150, and 200 µm for the 100 µm and 50 µm thick sheet, respectively. Each double row contains 30 holes which are interleaved (see figure 1). The sheets have been placed on the wedges such that the rows are oriented along the length of the wedges, i.e. every hole in a row is placed at a different wall thickness. Only the four larger diameters of each sheet have been used, i.e. 100–400 µm and 50–200 µm for the 100 µm and 50 µm sheets, respectively. A total of eight positions – 2 sheets, 2 wedges, 2 positions on a wedge – are sufficient to image every nominal hole diameter over any wedge thickness.

The assembled wedges are X-rayed using a microfocus setup at a magnification of 10. The microfocus tube ANDREX MX-4 was set to a voltage of $U=120$ kV and to a current of $I=800$ µA. The focal spot size has been determined using line pair gauges to approximately 70 µm full width at half maximum. The detector used was a Perkin Elmer XRD 0820AN digital detector array with a DRZ+ scintillating screen, a resolution of 1024×1024, and a pixel pitch of 200 µm. Every exposure covers an area of 2×2 cm, which corresponds to an effective pixel size of 20 µm. 20 frames were averaged per exposure for a total exposure time of 40 s. The raw frames have been corrected by nonlinear gain correction with 340 calibration frames at different tube currents, leading to a total calibration time of 11 minutes and a final signal-
to-noise ratio (SNR) between 400 and 600. In total 200 exposures depicting over 950 holes have been taken.

### 3.2 Image processing

Figure 2 displays one of the exposures. The radiograph is high-pass filtered for clarity only in this figure, for the experiment the image data have been left untouched. For the POD experiment with human operators the regular pattern of the defects is disturbing and leads to a much better visibility than images containing only single holes. Therefore smaller patches of 200×200 pixels have been cropped from the original exposures by an automatic algorithm such that a single hole is contained at a random position with a minimum distance of 25 pixels from the border. In order to determine the crop area, the position of the depicted holes must be known. This has been achieved by matching the exposures with previously performed high-precision measurements of the sheets at a synchrotron X-ray microscope with an accuracy of approximately 10 µm. The match of the high-precision model with the image data from this experiment was accurate to only 100 µm due to mechanical inaccuracy of the manipulation motors. Finally, the positions of the holes were known to an accuracy of 5 pixels. Additionally, more than twice as many patches guaranteed to contain no hole have been generated to serve as true negatives in the POD experiment. This is good practice for the POD data collection with humans. From all patches 319 images with holes and 645 empty images have been selected to be classified by human operators.

### 3.3 Collecting the experimental POD data

To collect the experimental POD data all of the 964 images had to be inspected by multiple human operators. As it presents a large effort and a significant burden to the operator to open
nearly 1000 image files and protocol the findings, a software tool has been developed to collect the data from mouse clicks, which aims to be very user friendly.

Figure 3 displays the main screen of the software tool as it is seen by the operator. The images are displayed in a row and using a scrollbar and arrows the user can page through the available images. The paging is smoothly animated to guide the eye of the operator. The user marks a defect simply by clicking on it with the mouse. A yellow cross indicates where the mark has been set. These marks can also by deleted using the mouse to correct an error. The contrast of the images can be enhanced by dragging a region-of-interest rectangle over the images. All user actions are automatically logged to an unbuffered file, which means that even in the case of a program crash the data is not lost. Three image sets are available: the first data set contains example images, the second one consists of a short series to practice the use of the software. The third data set is the aforementioned set of 964 images with the data for the POD trial. These image sets can be loaded by clicking on the appropriate buttons underneath the scrollbar. When the user chooses to finish the experiment a questionnaire is presented which tests for a number of inconveniences that might have influenced the result, e.g. fatigue, tedium, or problems with the usage of the software, and the level of experience with this kind of tasks. The experiment supervisor has access to a number of additional functions, which can be activated by pushing a hidden button, to automatically evaluate the data, plot the raw hit/miss data, and create a zipped file with the protocol collected during the run. Written instructions have been prepared to be given to the operator in printed form, which contain a simple overview of the task and a description of the usage of the software.

3.4 Results of the experimental POD trials

The task has been performed by five different operators with differing level of formal training and experience in RT. The fastest operator took only 36 minutes to complete the evaluation of the 964 images, while the slowest operator was busy for 1 ½ hour. The overall POD ranges from 48.3% to 64.9% and the false alarm rate is in the range of 0.4% to 12%. This indicates that the selected images are well chosen, as there are approximately as many hits as misses.

Figure 4 displays the experimental data and the POD($a$) curves determined by the software package mh1823 [8]. A logarithmic flaw size $a$ and a probit link was used in order to facilitate comparison with $a$-vs-$\hat{a}$ data from the simulations. The size $a$ was defined as the relative
change of wall thickness introduced by the defect. It is evident both from the raw data as well as from the critical thresholds, that the visibility of the holes improves with the diameter. For the holes with diameters of 300 µm and 400 µm the region where the POD approaches zero has not been reached, i.e. nearly all of the holes have been found by the operators. Therefore, no reliable conclusions can be drawn from the data. This also explains why the critical thresholds for the largest holes are slightly larger in this case, despite the general trend that larger indications are easier to detect. For the holes with a diameter of 100 µm and 200 µm the range of available wall thicknesses brackets the region where the POD changes most. Therefore, this data is suitable for comparison with the results from the simulations.

4 Simulated POD trials
Using the new POD facilities in aRTist the configuration has also been simulated. In order to facilitate the automatic evaluation, the geometric setup has been simplified and the source and detector specifications have been adjusted to match the experimental data.

4.1 The settings of the simulation
The setup consists of three parts. Along the central ray from the source to the detector the following objects are placed: a metal sheet made of pure titanium with a drilled hole, a thin wedge made of base material with the same inclination angle as the real wedge, and a cuboid.

Figure 4. Experimental PODs from the collected hit/miss data. The defects are cylindrical holes of the same length (100 µm) and four different diameters imaged over varying wall thickness. The symbols indicate the empirical POD from five human operators as a function of the wall thickness change introduced by the defects and the defect diameter. The solid black lines and dashed red lines show POD(α) and the associated 95% confidence bound, respectively.
of varying thickness, also made of base material. During a simulation run the thickness of the cuboid was increased linearly from 2.6 mm to 9.4 mm, which, together with the thickness of 0.5 mm of the wedge at the central point, yields a base material thickness corresponding to the range in the experiment. For every nominal diameter of the holes a Gaussian distribution was assigned to the diameter with the mean and standard deviation matching the precision measurement of the real holes.

The detector was characterised by its spectral sensitivity, which was calculated for the nominal thickness of the Gadox scintillating screen. The dose response and noise characteristics were specified to approximate the measured signal-to-noise ratio (SNR) over the whole range of grey values encountered in the experiment. Antialiasing has been performed using 20 samples per pixel. The sampling of the extended source focal spot was linked to the supersampling of the detector pixels.

The source spectrum was computed by aRTist’s Bremsstrahlung model for 100 kV with a 0.5 mm beryllium filter and a 1 mm aluminium filter. These settings were chosen to match the experimentally found attenuation of the base material. Unfortunately the voltage setting of the experiment (120 kV) did not result in a satisfying agreement. This may be partly due to the high voltage generator of the experimental equipment, which is unstable and poorly calibrated. We were also not able to accurately reproduce the differential gray value change induced by the hole sheet. The experimentally observed attenuation at the edge of the sheet corresponds only to a thickness of 65 µm, not 100 µm. This difference cannot be attributed to the scatter radiation from the object, since a quick estimate using Monte Carlo computations shows a scatter ratio of only a few percent for this configuration. In lack of a better understanding of this difference, we set the thickness of the sheet to 65 µm in the simulation in order to focus on the comparison of the POD calculation. With this setting the differential attenuation of the simulation closely matches the measured one over the whole range of wall thicknesses. Please note that these adjustments of the simulation, necessary to match the experiment, can be performed without actually manufacturing the artificial defects. Thus this

![Figure 5. Simulated $\bar{a}$-vs-$a$ data. The solid black line displays the linear regression, the 95% confidence bound is shown by the red dashed lines. The horizontal lines represent the signal threshold $a_{th}$, the decision threshold $a_{dec}$, and the saturation threshold $a_{sat}$.](image)
mismatch between experimental and simulated images is no obstacle in the practical application of the described simulation supported POD method.

4.2 Results of the simulation and comparison to the experimental POD

Figure 5 displays the simulated $\hat{a}$-vs-$a$ data for the 100 µm holes together with the regression and the thresholds for the POD($a$) computation. The signal $\hat{a}$ has been computed from the images using aRTist’s built-in model observer, which computes the statistical significance of a fit parameter associated with the flaw [7]. By visually inspecting the simulated images, the decision threshold $a_{dec}$ has been set to the point where the indication becomes unclear. The signal threshold $a_{th}$ was chosen such that clear outliers are excluded and the saturation threshold $a_{sat}$ was not used, i.e. set to a value outside the $\hat{a}$ data. The same procedure has been applied to all four diameters, except that for 300 µm and 400 µm the saturation threshold has been used to handle the nonlinearity. The settings and resulting POD parameters are displayed in table 1 together with the parameters from the experiment. A graphical representation is shown in figure 6.
Table 1. Results of the POD calculations

<table>
<thead>
<tr>
<th>Ø</th>
<th>Experiment</th>
<th>Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ø</td>
<td>µ</td>
</tr>
<tr>
<td>100</td>
<td>0.538</td>
<td>0.207</td>
</tr>
<tr>
<td>200</td>
<td>0.205</td>
<td>0.119</td>
</tr>
<tr>
<td>300</td>
<td>0.043</td>
<td>0.111</td>
</tr>
<tr>
<td>400</td>
<td>0.044</td>
<td>0.132</td>
</tr>
</tbody>
</table>

For the 100 µm holes there is a very nice agreement of the critical thresholds \( a₉₀ \) and \( a₉₀/₉₅ \) between the simulated and the experimental data. The noise standard deviation \( σ \) and the threshold \( a₉₀/₉₅ \) coincide within 1%, \( a₉₀ \) differs by 5%. This excellent agreement is not achieved for the larger hole diameters. Here the threshold \( a₉₀/₉₅ \) differs by up to 30% between experiment and simulation. However, the general trend that the larger holes at the same contrast can be found more easily is correctly shown by the simulation. The difference of the threshold \( a₉₀/₉₅ \) can partly be explained by the manual setting of the decision threshold during the \( a \)-vs-\( a \) analysis. Due to the construction of the POD(\( a \)) curve \( a₉₀ \) and \( a₉₀/₉₅ \) shift linearly with the decision threshold \( aₑₒ \). Hence, this setting, which is manually determined during the simulation, is critical for the result. A better point for comparison is the standard deviation \( σ \) or steepness of the POD(\( a \)) curve. Again, the good agreement for the smallest hole size is not found for the larger holes. A large part of this deviation can probably be explained by deficiencies of the experiment. Since the hit/miss data is centered on the threshold only for the smallest holes, the experimental data is unreliable for the larger holes. Even for a diameter of 200 µm the region where the holes are invisible is barely contained in the data (cf. figure 4). On the other hand, the simulated \( σ \) scales inversely proportional to the hole diameter, which is very reasonable in the context of classical models for visibility [9,10], but not reflected in the experimental \( σ \). This indicates that there is not enough data for all but the smallest holes and the comparison is therefore only reliable for the latter.

The overall simulation supported POD methodology employed in this paper can be outlined with the following steps. First, a sheet metal of the same material as the test piece must be obtained. Second, a few images must be taken of the sheet edge placed at different positions on the test piece or a dummy wedge with varying thickness of the base material. Third, the simulation is adjusted to match the experimentally found grey values, differential attenuation, signal-to-noise ratio, and unsharpness. Fourth, a batch of images is simulated and the defect visibility is computed using a statistical model observer. Finally, the decision threshold is set by inspecting the simulated images. Then, the POD(\( a \)) curve is computed using standard analysis procedures. At least, this process can be applied to digital radiography, where the images are evaluated on a computer screen. For film radiography, the evaluation of the film on a lightbox without the means of digital contrast enhancement imposes additional restrictions on the visibility, which are not reflected in the method.
5 Conclusion
With the help of wedges and sheet metal with drilled holes, the POD of pores in a microradiograph has been measured experimentally. For this purpose, a software tool has been developed to collect many samples easily. The whole configuration has been simulated successfully using aRTist including the image evaluation by the human operator. For small indications with good experimental support, an excellent agreement of the POD within a few percent has been found. For larger indications, good qualitative agreement is observed. Adjustments have been done to the simulation software to closely match the experimental data, which readily lead to a simulation supported POD methodology.

Acknowledgements
We thank Ralf Britzke, Axel Lange, Bernd Müller, and Thomas Wolk for their support with the measurements at BESSY. We are grateful to Uwe Zscherpel for valuable discussions and hints. Financial support by the European commission in the project PICASSO under grant 234117 (FP7-AAT-200-RTD-1) is gratefully acknowledged.

References