WHAT HAVE WE LEARNT FROM THE EVALUATION OF AN RRT?

F. Fücsök[1], C. Müller[2], M. Scharmach[2], M. Elaguine[2]
[1] Budapest Power Plant Ltd, Budapest, Hungary, [2] BAM-Berlin, Berlin, Germany

Abstract: In order to estimate the performance of representative inspectors of industrial NDT laboratories a round robin test was organized in Croatia, Hungary and Poland. Altogether 60 results were evaluated.
The evaluation method of the round robin test is the receiver operation characteristic, in short form ROC. A short introduction of the ROC method with an example will be presented in the paper.
Discussing the circumstances of the original RRT where all visible discontinuities had to be recorded, many of the participants stated that they did not exactly follow the prescription of the test. They had thought that the small discontinuities, e.g. below 3 mm diameter, could be omitted because even the strictest acceptance level allows them. The most practised evaluators did not write the small gas and slag inclusions on their list of flaws detected, the trial then yielding an incorrectly low figure for their reliability of detection. Consequently, two types of evaluation were made: one taking into consideration all of the discontinuities and the other only the bigger ones that were over acceptance level 1 of EN 12517:1998. We have learnt that we had to follow the real circumstances of the evaluation.
The performance of the inspectors was measured with the maximum operation points, i.e. each inspector's maximum POD and their PFA at that point. We consider this to be more realistic than other possibilities such as the area under the ROC curve or *K* value of the curve.
It was surprising that as many as one-third of the ROC curves have an S form. Statistical analyses taught us that the cause of the S form of ROC curves are different forms of distributions, not following the simple model of two Gaussians, one for signal distribution and one for signal+noise distribution.

Introduction: The reliability of a diagnostic system is an important question for the human doctors, for the fracture mechanic experts, or for the customers of NDT laboratories. But measuring the reliability is a difficult problem because it depends on a lot of elements. A good method for laboratories to demonstrate the reliability of its work is to participate in round-robin tests.
If you are in a fieldwork you will find that everybody can evaluate the radiographic films. So the most serious quarrels are about the results, with other words the reliability of the radiographic test. That was the reason for the topic of an international RRT, the radiographic film evaluation. The RRT was organised for voluntary industrial laboratories in Croatia, Hungary and Poland. Altogether 60 participants from 22 laboratories took part in the test.
According to the modular approach it is possible to analyse the reliability of a system by separate modules. So we can use the results of the reliability measurements of radiographic film evaluation like one module of radiographic testing.
For the purpose of the RRT a set of films was selected in the Reliability Laboratory of BAM, which provides the scientific and technical support. The selected 38 films contain 206 defect indications of different types and dimensions.
The selected films were scanned by a state of the art film digitiser. Removing the unnecessary areas the X-ray films were copied with a laser printer. The digital images were evaluated with the help of a dedicated image processing computer program of BAM. Later the types of the discontinuities were discussed and after some changes agreed on by a small group of Hungarian experts. The results of this evaluation are taken as the true values of the discontinuities.
Four sets of films were printed into AGFA Scopix Laser films, which were sponsored by AGFA. In this way all of the participants of the RRT evaluated exactly the same discontinuities on equivalent films. During the test there were not any complain about the quality of copies, despite

the fact, that the films were copied what we wrote to the participants. We learned that the digital copies were good enough for correct evaluations.
The voluntary evaluators were provided with clear instructions of the procedure and specific forms for the support of the evaluation work and to aid the pre-processing of the results with the computer. The forms contain the following columns: Identification, the code of imperfections according to EN 26520:1991 standard, the co-ordinates, the dimensions and the detectability of the discontinuities and for completeness the IQI and the optical density of the films.
The inspectors were asked to evaluate and identify each weld image cm by cm, which is a very strict prescription. Additionally they were required to fill in a form for the circumstances of the film evaluation including the length of evaluation time. They stated that the evaluation work of the 38 films took 5 to 12 hours. The longer evaluation time the more detailed discontinuities were indicated.
The evaluation results were filled in an Excel table to support the data processing of the indication results for the ROC statistics.
During the time of RRT we asked the opinions of the participants about the test, many of them expressed, they could not exactly follow the guidelines of the test. According to their professional experience the small discontinuities could be omitted because even the strictest acceptance levels allowed for them. So the most experienced evaluators did not write the small gases and slag into the list, which caused an incorrect reliability of detection. As a consequence we should accomplish two types of evaluation. In the first we took into account all discontinuities and in the second only the bigger ones, which were over the acceptance level 1 of the EN 12517:1998 standard.
For evaluation of the reliability, we determined the Receiver Operating Characteristic (ROC) curves. The full ROC curve was determined from four points representing the different detectability of discontinuity, which where selected by the evaluators. The curves were calculated by maximum likelihood fit.

Results: On Figure 1 you can see an ROC curve, which was given to the participants as feedback after the test. The figure contains three curves. The higher curve is the result of the best evaluator, the lower one is the worst result. Between them you can found the result of the participant who was coded by the number in the legend. The most interesting point is the highest point of the curve, which contains the probability of all recognised discontinuities, so we called it the working point.

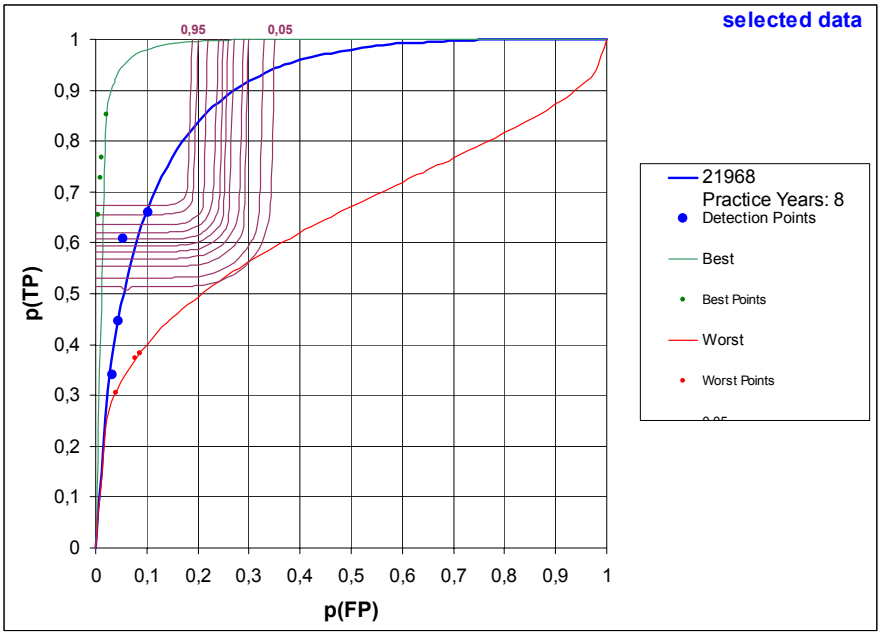

Figure 1. The results of the ROC evaluation of selected data

In the left upper corner of the diagram you can see a set of curves. They represent the requirements of the ASME code XI. Appendix VIII. The curves show the 5%, 10%....80%, 90% and 95% reliability level or probability of acceptance. The working point can be found between the 90% and 95% reliability level. It means, the evaluator will met the ASME Code requirements with more than 90% probability.
We used the ASME Code requirements arbitrary, but written requirements for probability of detection can be found only for the evaluation of ultrasonic testers.
The evaluation of the participant mentioned above was made by the results, which take into account discontinuities, which are over the acceptance level 1 of the EN 12517:1998 standard. At Figure 3 the ROC of the same participant can be seen, where all data were evaluated. It can be recognised the reliability of evaluation of failures is better than the evaluation of all discontinuities.

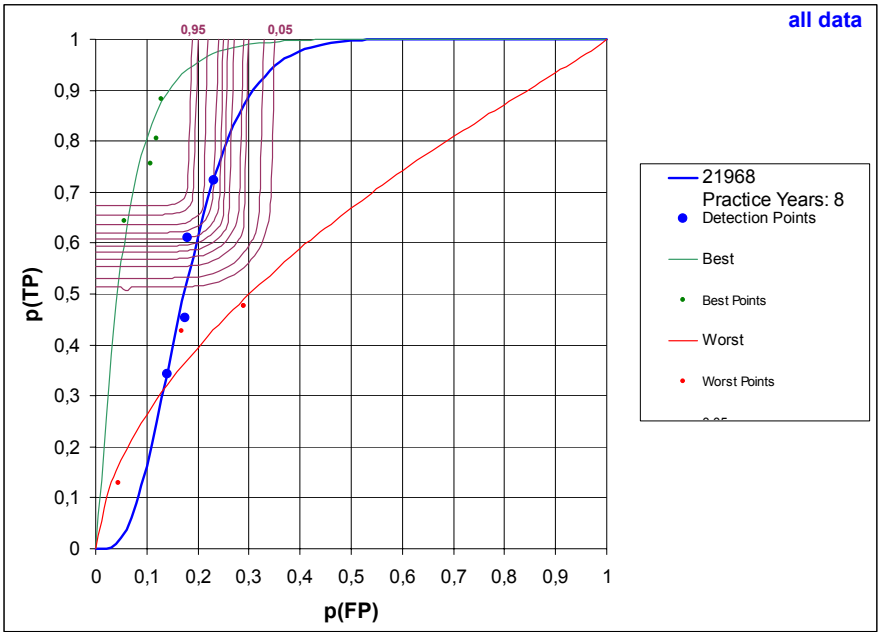

Figure 2. The results of the ROC evaluation of all data

The best and the worst evaluator were chosen by the sequence of distance of working points from the diagonal: $k_{wp.}$ The interpretation of k_{wp} can be seen at the Figure 3. We found this distance to represent the level of reliability better than the K vector, which was used before, or the area under the curves.

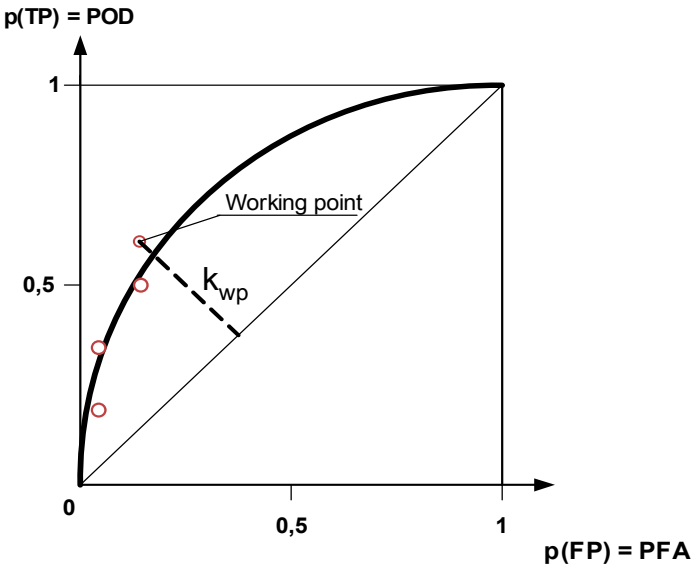

Figure 3. The interpretation of distance of working point

Figure 4 represents the sequence of participants when they evaluate failures i.e. unacceptable defects. Figure 5 represents the results of evaluation all of the discontinuities.
The distances of working points of the second evaluation method are smaller, which document the evaluators' good practice to work according to the standard.

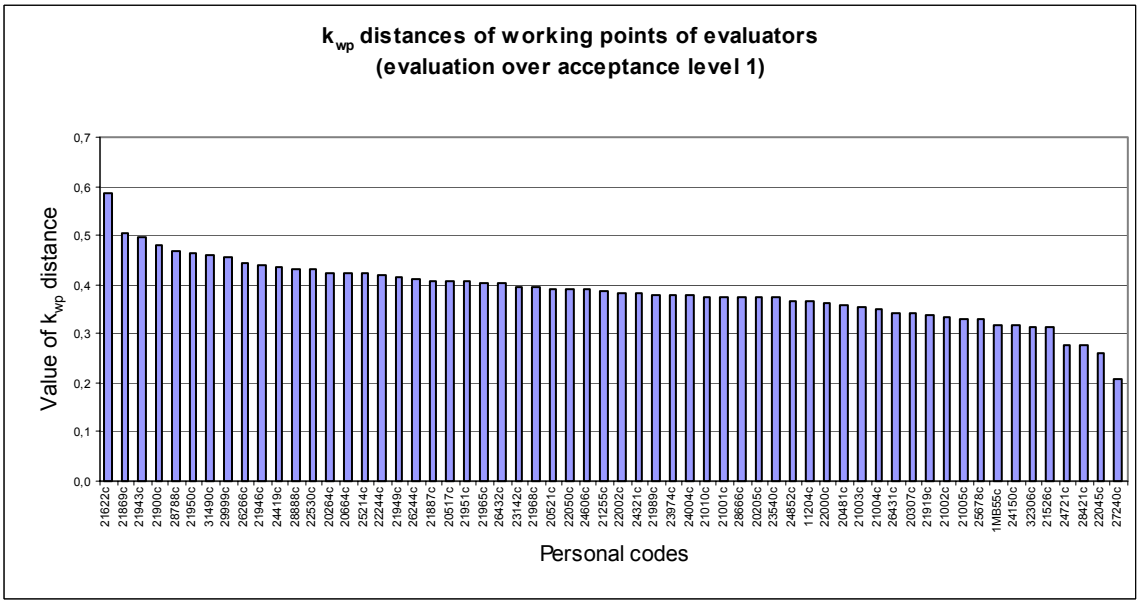

Figure 4. The sequence of participants of failures evaluation

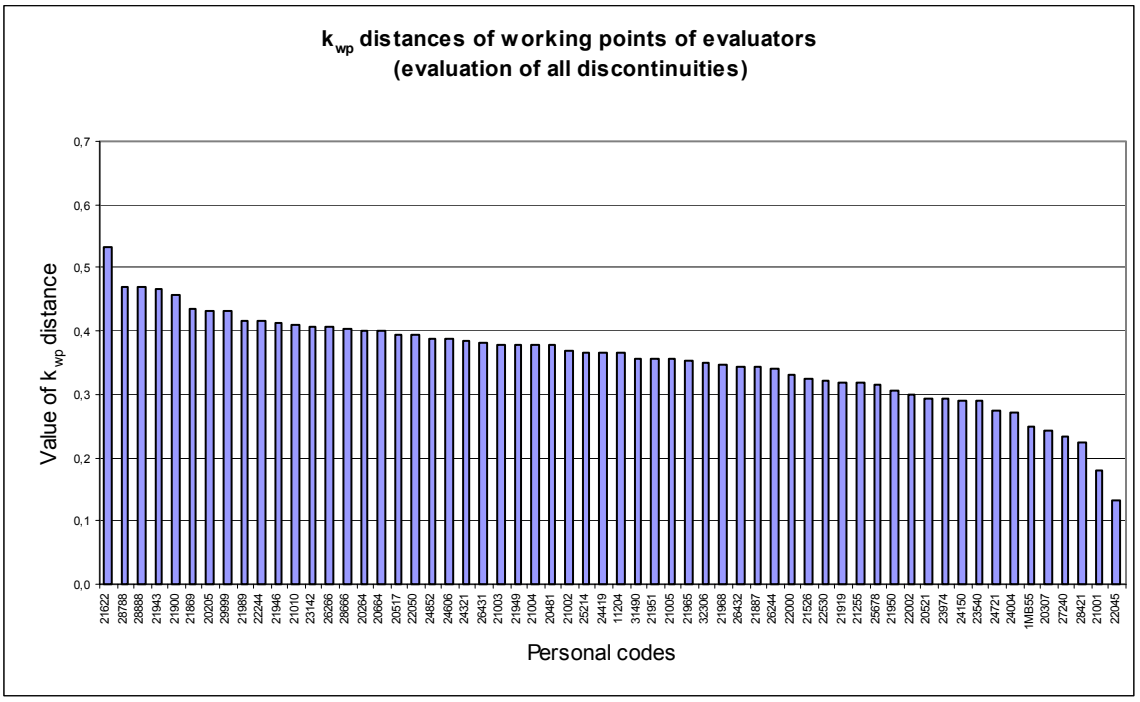

Figure 5. The sequence of participants of all discontinuities evaluation

During the evaluation we tried to find the correlation between the years of experience of the participants and the probability of detection. Naturally there were differences of performance of the older and the younger tester, but the differences were not significant.

Discussion: First of all we had to learn that we tested the human factor of NDT. Testing the human factor you can find a lot of problems with the "tested elements". The most interesting was, that the tested person knew that he was under a test. In our round-robin test we supposed that we could model a long boring workday. But it became clear that the evaluators were able to concentrate all day, and in this point of view it was unnecessary to choose more than 200 inflections for the evaluation. Otherwise, almost half of the participants omitted the small discontinuities. In this reason, as was mentioned, we had to make and evaluate two sets of data.
Another problem was some misunderstanding of the filling the tables. Some defect indications had to be corrected, but this action may modify the results of the participants, which is forbidden in the theory of measuring. On the other hand the evaluation of RRT was made by computer program. It can't evaluate the strange data format, so we had to correct the wrong formats of data. But sometimes we found data errors, which were evidently simple wrong counting or other misunderstanding. These types of errors were not corrected because the correction would modify the reliability results of participants. We did corrections only for those cases they did not modify the results of reliability.
The second set of problems is the true values. Till the end of the round-robin test we won't publish the true values. But the participants are deeply interested in this particular element, because they are only familiar with this area and they know that the result of round-robin test depends on the true values. But, it is well known that one cannot determine exactly the true values.
In a welded joint numerous combinations of inflections are possible. So some different but mainly equivalent solutions can be determined in its evaluation. It is certain, another expert will refuse and modify the earlier evaluation. If you want to eliminate the disagreement in the RRT you have to delete the most discussed radiographic films. In this way you will withdraw a little

from the real life. So, you will not determine the true values correctly, consequently you could not specify the reliability of the evaluation with 100% reliability.
Theoretically all of the measurements have some uncertainty. In the planning of a round-robin test of non-destructive testing you have to focus your attention on the determination of true values.
You have to choose, as best you can, real discontinuities and the determination of true values have to be as correct as possible. If you use destructive methods to determine the true values they will cost a lot of additional money. But it is not certain it will be more accurate than the non-destructive methods.
We have to learn too, you must choose simple characteristic to represent the reliability performance because the NDT technicians will not understand them when they are too sophisticated.
One of the simplest characteristics that can be used is the so-called K vector, which is easy to understand, and it is good for representing the performance of the evaluators. It was suggested by Van Dijk and Boogaard. [1]
But we recognised the results of evaluation with K vector were not correct. Searching for the reasons we found a lot of ROC curves which were not hyperbolic. Counting the S shapes ROC curves we found 20 pieces among all data evaluation and 13 pieces among the failures evaluation. Statistical models represent the form of ROC curves depend on the distribution of the noise and the noise+signal.

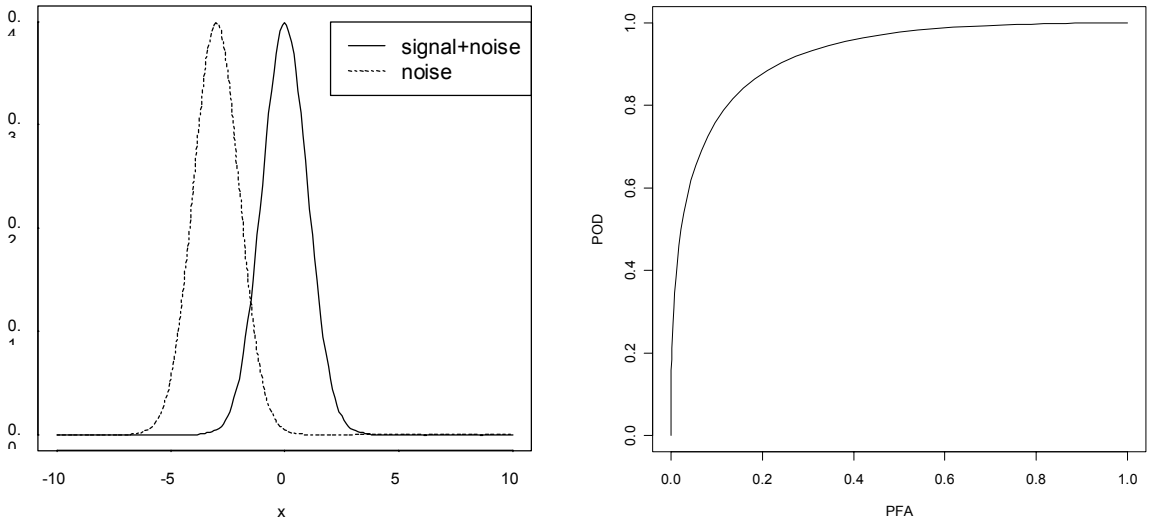

Figure 6. a The hyperbolic form of ROC curve

If the distribution of noise and noise+signal are approximately the same the form of ROC curve will be hyperbolic. In figure *6.a* a simple model that we always use is presented. But, if the distribution of noise becomes wider and lower compared to the noise+signal's distribution the form of ROC will change, as you can see in the figure *6.b*. After having checked the results once more we found one curve, which had this strange form. It can be seen at figure 1 as the curve of the worst result.

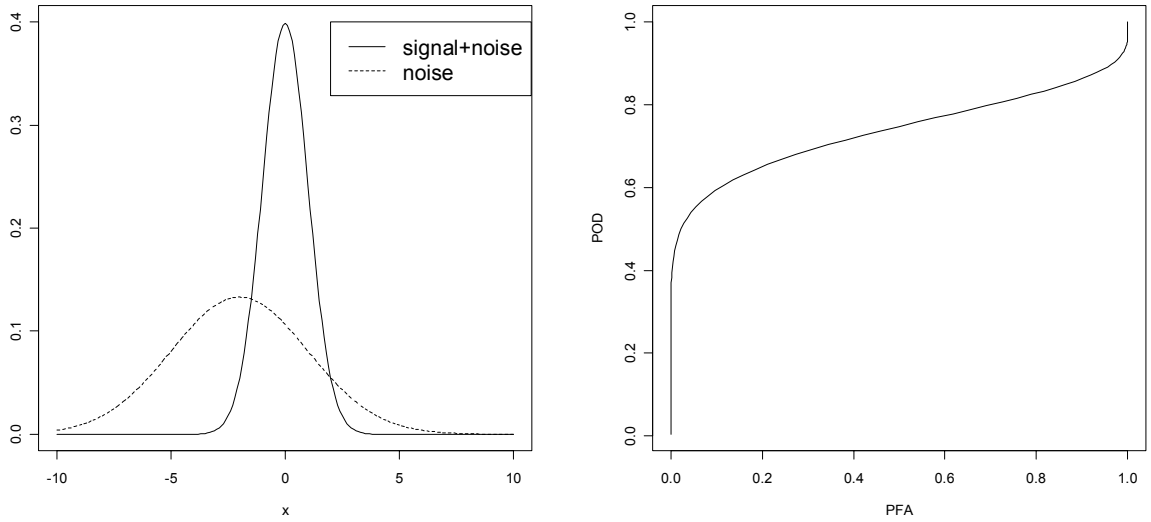

Figure 6. b The strange form of ROC curve

If the distribution of signal+noise changes into wider and lower compared to the noise's distribution the form of ROC curve will change into S form. The form of ROC can be seen in the figure *6.c*.

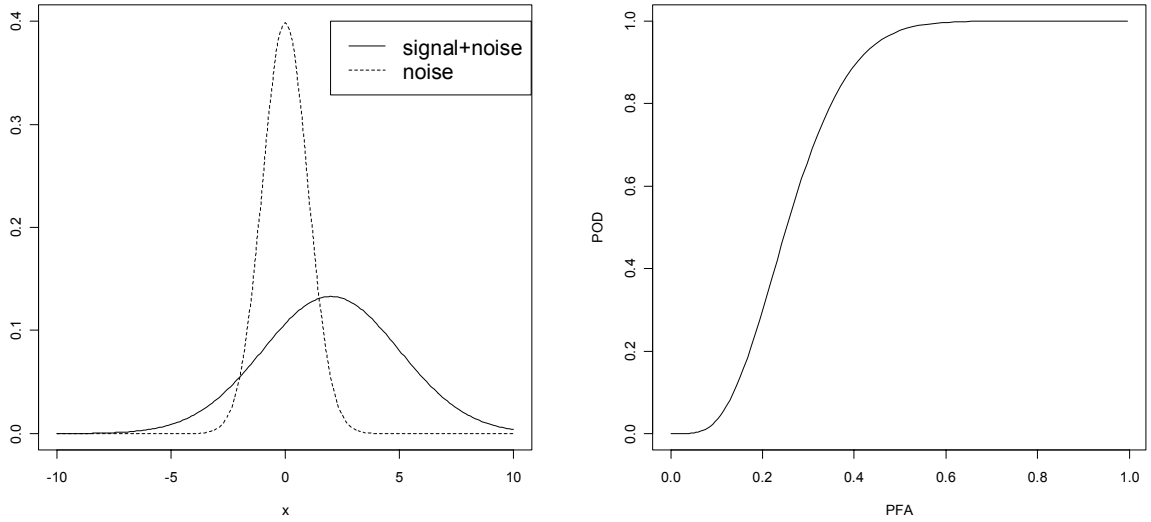

Figure 6. c The S form of ROC curve

This statistical evaluation helps us to understand why the K vector of ROC curves was not able to compare the performance of participants. So the distance of working point from the diagonal is better to compare the performance, as we used it.

Conclusions: We summarized the place of working points among the reliability levels. In table 1 the reliability results of participants can be seen. In the first row you can see the domains of the reliabilities to fulfil of the demands of ASME Code. In the second row can be seen the number of persons having a working point in the certain domain in the first row, evaluating all discontinuities. In the third raw can be seen the number of persons having the working point in the certain domain in the first row, when the failures were evaluated.

Table 1. The reliability results of participants to fulfil of the demands of ASME Code

Probability domains [%]	<5	5 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	61 - 70	71 - 80	81 - 90	91 - 95	**> 95**
Persons when evaluate all discontinuity	7	2	3	2	1	1	3	7	5	9	5	**15**
Persons when evaluate the failures	4	1	3	3	2	0	1	0	14	4	7	**21**

The number in the last column shows that 15 persons fulfil the demands by higher than 95% reliability when all the discontinuities were taken into consideration. In the last row you can see that 21 person, more than 1/3 of participants, fulfil the demands by a higher than 95% reliability when the failures were evaluated.

The authors' mean it is currently not their task to evaluate if these results are enough for the fracture mechanic experts, or for the customers of NDT laboratories, but might be within a future risk based management. In any case it is evident, the reliability of radiographic film evaluation should be developed.

References: [1] Van Dijk at al., Non-Destructive Testing 92, C.Hallai and P. Kulcsar (Editors), 1992 Elsevier Science Publishers B.V., pp xxxi