Calculating Inspector Probability of Detection Using Performance Demonstration Program Pass Rates

Stephen Cumblidge[1, a)] and Amy D'Agostino [1, b)]

[1] United States Nuclear Regulatory Commission, Washington, DC 20555-0001

[a)]Stephen.Cumblidge@nrc.gov
[b)] Amy.DAgostino@nrc.gov

Abstract. The United States Nuclear Regulatory Commission (NRC) staff has been working since the 1970's to ensure that nondestructive testing performed on nuclear power plants in the United States will provide reasonable assurance of structural integrity of the nuclear power plant components. One tool used by the NRC has been the development and implementation of the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code Section XI Appendix VIII[1]) (Appendix VIII) blind testing requirements for ultrasonic procedures, equipment, and personnel. Some concerns have been raised, over the years, by the relatively low pass rates for the Appendix VIII qualification testing. The NRC staff has applied statistical tools and simulations to determine the expected probability of detection (POD) for ultrasonic examinations under ideal conditions based on the pass rates for the Appendix VIII qualification tests for the ultrasonic testing personnel. This work was primarily performed to answer three questions. First, given a test design and pass rate, what is the expected overall POD for inspectors? Second, can we calculate the probability of detection for flaws of different sizes using this information? Finally, if a previously qualified inspector fails a requalification test, does this call their earlier inspections into question? The calculations have shown that one can expect good performance from inspectors who have passed appendix VIII testing in a laboratory-like environment, and the requalification pass rates show that the inspectors have maintained their skills between tests.

While these calculations showed that the PODs for the ultrasonic inspections are very good under laboratory conditions, the field inspections are conducted in a very different environment. The NRC staff has initiated a project to systematically analyze the human factors differences between qualification testing and field examinations. This work will be used to evaluate and prioritize potential human factors issues that may degrade performance in the field.

INTRODUCTION

This work was motivated by the NRC staff need to assess the effectiveness of the ultrasonic testing (UT) performed on nuclear power plant welds by inspectors using UT. It is impossible, however, to directly measure effectiveness of UT in the field, as we do not know the true state of the components being examined in nuclear power plants and the relatively few flaws present in nuclear power plant welds would not allow for a statistically-significant measurement of the inspector capabilities. As direct measurements are impossible, indirect measurements in the laboratory were used to estimate the probability of detection for inspectors in a laboratory environment, followed by a planned assessment of the differences between laboratory and field examinations.

The primary data for this work comes from the Performance Demonstration Initiative (PDI). PDI is the primary administrator for the ASME Code Section XI Appendix VIII blind qualification tests for ultrasonic inspectors in the United States. The Appendix VIII tests are broken into a series of supplements, based on the materials and

geometry of the component to be examined. Examples include austenitic (stainless) steel welds, which are covered under Supplement 2, ferritic steel welds are covered in Supplement 3, and dissimilar metal welds (between austenitic and ferritic components) are covered in Supplement 10.

A special set of austenitic supplement 2 stainless steel welds were harvested from cracked boiling water reactors in the 1980s, and contain intergranular stress corrosion cracks (IGSCC) that developed during the operation of the power plants. These specimens are covered under an examination titled "Supplement 2 with IGSCC." This Appendix VIII qualification is unique in that it needs to be repeated and passed every three years to maintain the qualification.

PDI has conducted thousands of tests on a large number of inspectors over the past twenty years. The NRC has been meeting with PDI regularly to monitor the development and application of the tests to assure that the PDI tests meet the requirements of ASME Code Section XI Appendix VIII. It is worth noting that Appendix VIII was primarily designed as a screening tool to weed out ineffective procedures, equipment, and personnel, and was not designed to provide a precise measure of the POD for the inspectors. Additionally, PDI was set up to administer a set of pass/fail qualification tests, and the PDI staff did not design their testing protocols to collect and quantify the testing data in a scientific manner for later statistical analysis. The information has been recorded and is currently being entered into a robust database, but this work is ongoing and the data available for this paper was somewhat limited.

While the information at PDI is not complete, the PDI staff have worked with the NRC staff on select cases where the PDI data was re-examined to allow for complex analysis of some data sets. During the meetings between NRC and PDI staff, a common statistic that PDI has been able to provide is the number of people taking the qualification tests and the number of people passing the PDI tests.

Over the years, pass rates for some PDI qualification tests have been close to 50%, which is similar to reported pass rates for international performance demonstration programs. In addition, the requalification testing for IGSCC are roughly equivalent to the initial pass rates. It was initially expected by the NRC staff and industry that the pass rates would increase after the weaker inspectors were not qualified. Both of these realities have raised questions and concerns.

This analysis was performed to try to answer or provide insight into the following questions:

1. How skilled are nuclear NDE inspectors in general?
2. Are large numbers of poorly-qualified inspectors taking the test and passing via luck?
3. Why are requalification pass rates for IGSCC testing similar to the pass rates for the initial tests?
   a. Does the requalification rate raise a safety concern?
   b. Does a failed requalification test call previous inspections into question?

PROBABILITY OF DETECTION CALCULATIONS

ASME Code Section XI Appendix VIII Design

ASME Code Section XI Appendix VIII is used in the US to demonstrate the capabilities of ultrasonic procedures, equipment, and personnel for use in a variety of ultrasonic inspections. Appendix VIII contains requirements for performance demonstrations for nine different materials and components, including welds in austenitic piping, welds in ferritic piping, pressure vessel inspections, and bolting inspections.

The goals of the Appendix VIII tests are to ensure that inspectors with an unacceptable probability of detection would have a low probability of passing the test while skilled inspectors should have a good probability of passing the tests. The ideal way to do this would be to have each inspector look at a large series of specimens to evaluate the inspector POD for a series of flaw sizes to precisely measure each inspector's skill. This approach is, unfortunately, prohibitively expensive and time consuming. As a large series of tests was not feasible, a smaller "weed out" test was designed using binomial power curves. The pass/fail criteria were determined for several sample sizes to set a minimum chance of passing for different inspector skill levels, taking statistical uncertainties inherent in small sample sizes into account. As a result, tests with the fewest possible flaws are significantly more difficult than those with the maximum possible number of flaws to prevent an unskilled inspector from passing via luck. Another controlling factor is that all tests must give an inspector with a low POD a very low chance of passing the test. The 5/5 test is significantly more difficult than a 14/20 test, but it was decided during the design of Appendix VIII that a

4/5 test would allow poor inspectors through too often. The power curve equation is shown in Equation 1 and example power curves for different tests are provided in Figure 1

$$Pwr(POD) = \sum_{k=C_1}^{M} \left(\frac{M!}{k!(M-k)!} \right) POD^k (1-POD)^{M-k} \qquad (1)$$

Where:

POD	=	Inspector Probability of Detection
Pwr(POD)	=	Probability of passing a test with a given POD
M	=	Total number of cracked weld units inspected
C1	=	Threshold for POD test.

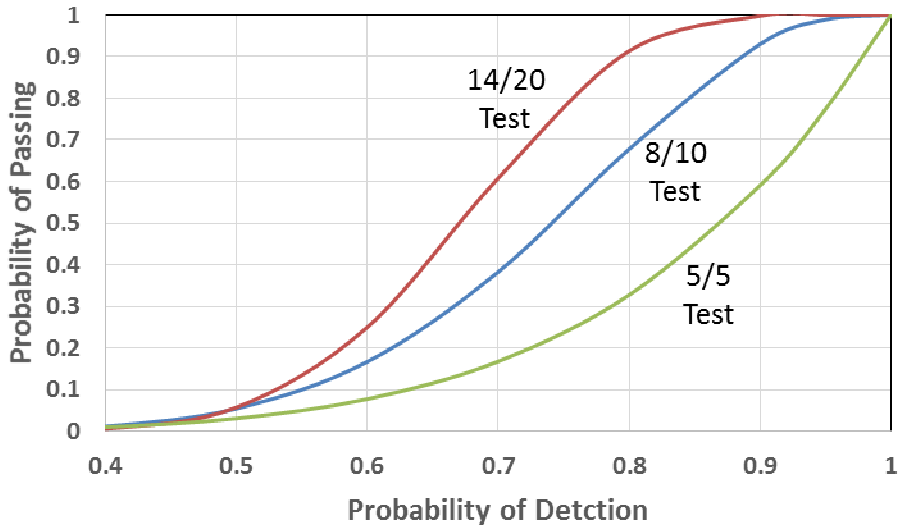


FIGURE 1: Example Power Curves for three possible tests, one for a test with 20 flaws, of which 14 must be found to pass, one for a test with ten flaws, eight of which must be found to pass, and one for a test with five flaws, all of which must be found to pass.

Similar power curves were used to develop the pass and fail criteria for false calls. Controlling for false calls is important, as, an ineffective procedure can pass a pure detection test by luck. The use of such a procedure could result in unnecessary repairs in the field.

Pass Rates for Appendix VIII Testing

The primary data for the following calculations comes from the performance demonstration testing performed at PDI to meet the requirements of Appendix VIII. PDI administers the Appendix VIII testing for US licensees and vendors for the various tests required in Appendix VIII. The NRC and PDI staff meet annually to discuss any changes or developments in the Appendix VIII testing. One area that is discussed at the NRC/PDI meetings are the total number of people taking the PDI tests and the fraction of those that pass. The data presented in this paper was taken from the January 2015 meeting between the NRC and PDI [2] and includes data from 2011 to near the end of 2014.

PDI allows inspectors three tries per year to pass the Appendix VIII tests. The ability to retest up to twice was not part of the original design of Appendix VIII, but is not prohibited. The PDI test results data is thus given as the

initial number of inspectors, the numbers passing the first, second and third tests, and the final yield after three tries. Not all inspectors who fail the first and second test retry the later test, thus, the yield after three tests is a combination of the inspectors who failed and those who chose not to retest. As a point of note, PDI separated the testing results for encoded and non-encoded inspection methods at the request of the NRC staff. Table 1 contains the numbers of inspectors attempting and passing each of the three possible tests, the pass rate for the first attempt, and the yield after all three attempts for manual (non-encoded) ultrasonic testing procedures and Table 2 has the same information for encoded ultrasonic procedures. It is worth noting that the "Ferritic" test cited was given as an add-on test to the Austenitic test and is given as a 5/5 test. Figure 2 shows the initial pass rates for each technique with the associated 95% error bars (90% inclusive) for manual (non-encoded) and encoded examinations.

Table 1: Numbers of inspectors taking and passing each of the three possible attempts for each type of PDI tests for six detection qualification tests using manual (non-encoded) ultrasonic inspection techniques

Qualification Test	1st Attempt		2nd Attempt		3rd Attempt		Initial Pass %	Total Yield %
	Take	Pass	Take	Pass	Take	Pass		
Austenitic with no IGSCC	60	29	22	15	5	4	48	80
Austenitic with IGSCC	65	41	18	10	3	2	63	82
IGSCC Requalification	110	63	43	18	11	10	57	83
Ferritic (5/5 test)	125	62	43	30	9	9	50	81
Weld Overlay	44	32	9	6	0	0	73	86
Dissimilar Metal Weld	16	13	3	3	0	0	81	100

Table 2: Numbers of inspectors taking and passing each of the three possible attempts the PDI tests for four detection qualification tests using encoded ultrasonic inspection techniques

Qualification Test	1st Attempt		2nd Attempt		3rd Attempt		Initial Pass %	Total Yield %
	Take	Pass	Take	Pass	Take	Pass		
Austenitic with IGSCC	17	14	1	0	0	0	82	82
IGSCC Requalification	24	12	8	7	0	0	50	79
Ferritic (5/5 test)	11	8	0	0	0	0	73	73
Dissimilar Metal Weld	19	14	5	4	0	0	74	95

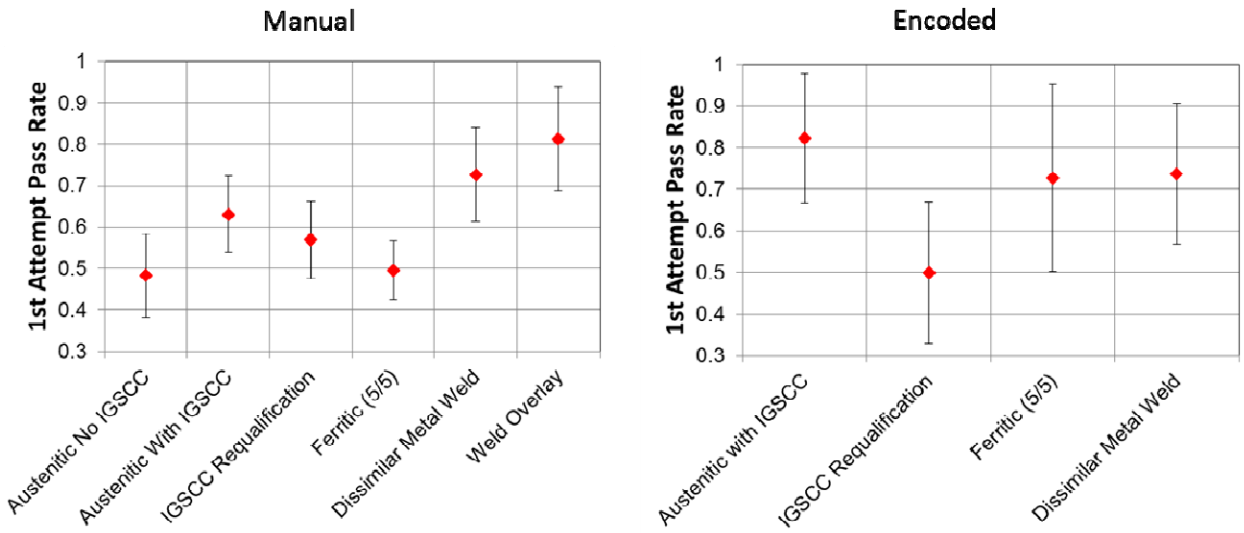


FIGURE 2: Pass rates of 1st attempt for non-encoded and encoded ultrasonic qualification tests

What is interesting, and has been a subject of discussion for several years between the NRC and PDI staff, is that the IGSCC requalification pass rates are slightly lower than the first time pass rates. For the non-encoded examinations they are within the level of statistical uncertainty, but for the encoded examinations there is a measurable difference between initial qualification success rate and the requalification success rate. The IGSCC requalification rate with encoded inspections is within the statistical uncertainty of the manual initial qualification.

While this information is useful, it has limitations. The PDI staff were not able to provide the false call rates for the different qualification tests or the number of inspectors who failed the tests because of false calls. Additionally, it was not recorded why people declined to re-take tests after they failed.

Estimating Inspector POD using Pass Rates

Given the power curves used to design the Appendix VIII tests and the pass rates for the various tests, back calculating the average inspector POD using the pass rates appears to be relatively simple. While the power curves may show that if one runs 100 inspectors with a POD of 74% though an 8/10 test that 50% of them will pass. This does not mean that if you run 100 inspectors through the same test and they have a 50% pass rate that they have, on average, a 74% POD. The pass rate information alone does not tell you the initial skill distribution of the inspectors taking the test. A 50% pass rate could mean anything from the extreme case that 50% of the inspectors have perfect performance (100% POD and no false calls) while 50% are incapable of finding any flaw (0% POD) to the case where all inspectors have a POD of exactly 74%. The pass rate does give an insight into the minimum skill level of those who pass the test however. The skill level of those who pass will range from 74% to 100%. Fortunately, the PDI testing allows for up to two retests, providing the total yield after three tries. If the initial distribution was at the extreme level, with an even mix of 100% and 0% POD inspectors, the yield after three tries would be identical to the initial pass rate, i.e. 50%. If he inspectors had all had a 74% POD, the yield after three tries would be 87.5%. These cases are explored in Figure 3.

Consequently, using the pass rate to back-calculate the inspector POD using the power curves yields the worst possible average inspector skill for those who pass the test and can be considered conservative. As the pass rate information from PDI does not differentiate between inspectors failing because of false calls, this value is doubly conservative. Also, looking at the PDI pass rates after three tries, which are higher than the pass rate for the first try but lower than 1-(initial pass rate)^3, we can see that the initial distribution is apparently somewhere between these extremes.

One issue with using the power curve information to estimate the POD for the inspectors is which allowable power curve to use for the estimation, as some tests can be conducted using anywhere from five to twenty flaws. The 14/20 test has a very different profile than the 5/5 test and would give very different results. For this work the 8/10 test was chosen as the default when the actual value was not known. The 8/10 test was chosen as some qualification tests require at least ten flaws, and the 8/10 test is a “medium” difficulty test. Figure 4 shows the lowest estimated POD for field inspectors for non-encoded and encoded tests using the PDI pass rates, assuming an 8/10 test, except for the ferritic test which were conducted with a 5/5 test.

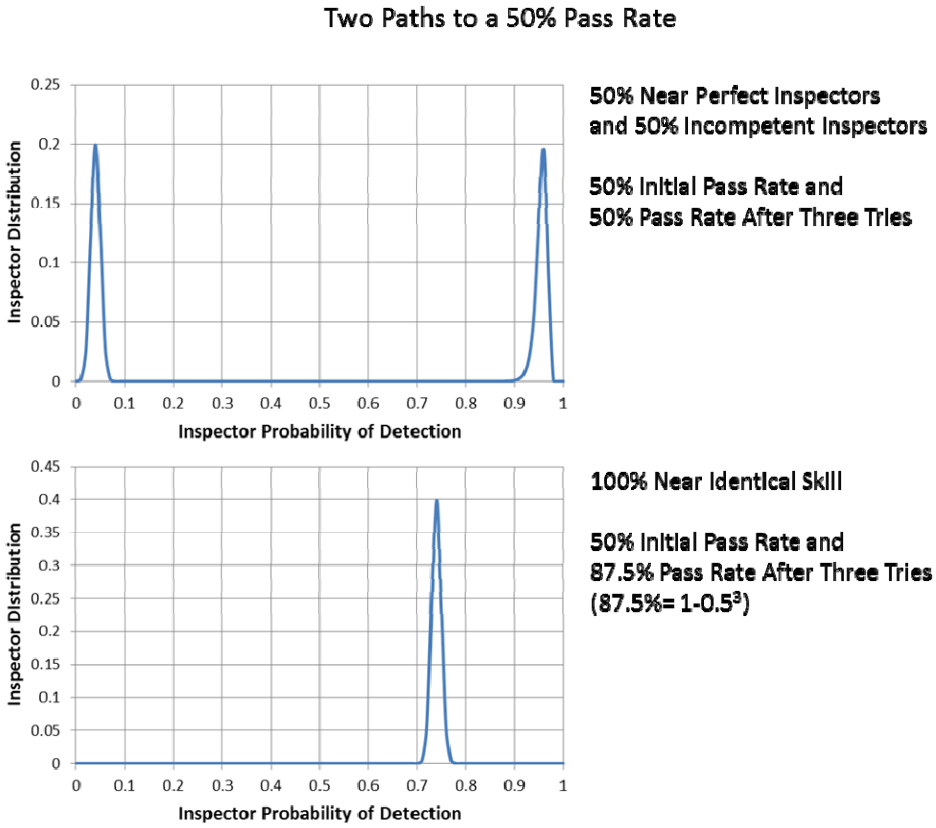


FIGURE 3: Postulated Extreme Initial Inspector Skill Distributions and Their Effects on Initial Pass Rates and Total Pass Rates after Three Tries

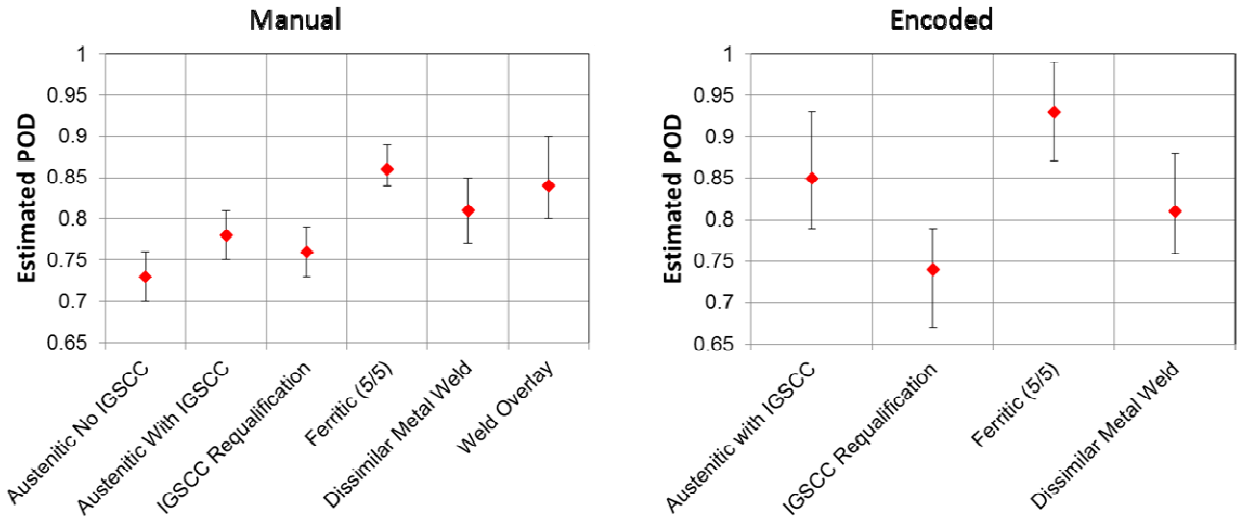


FIGURE 4: Estimated conservative inspector PODs for non-encoded and encoded ultrasonic qualification tests

Evaluating the Effects of Qualification Testing on Inspector Skill Distributions

While we do not know the skill distribution for the inspectors taking each qualification test, we do know the initial pass rate and the yield after three attempts to take the tests. With this information and the power curve rules, one can produce a hypothetical distribution for the POD of a set of inspectors, multiply this distribution with the power curve for a given test, and then tally up the fraction that pass and fail in two new distributions. This method is illustrated in Figure 5. Using this method one can try to get an idea how the testing would skew sets of hypothetical inspector skill distributions to test different cases and to estimate the level of conservatism reflected in the worst case PODs shown in Figure 4.

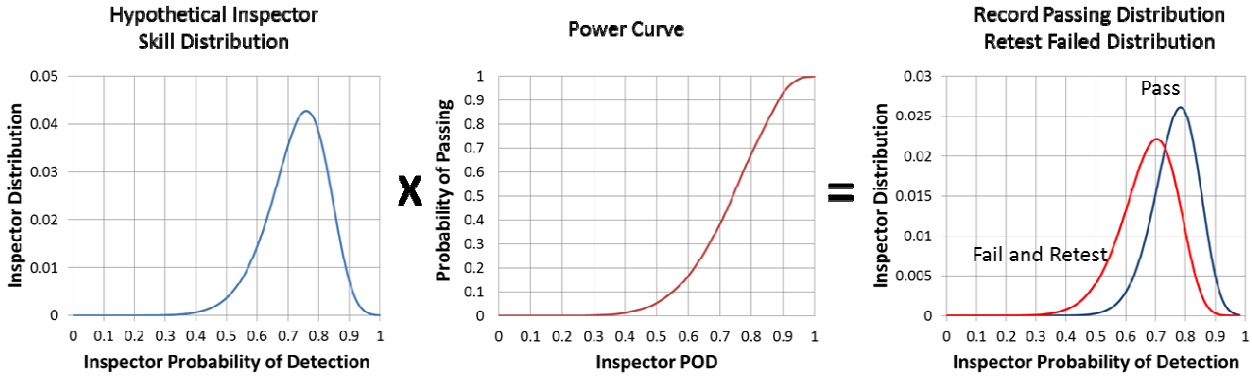
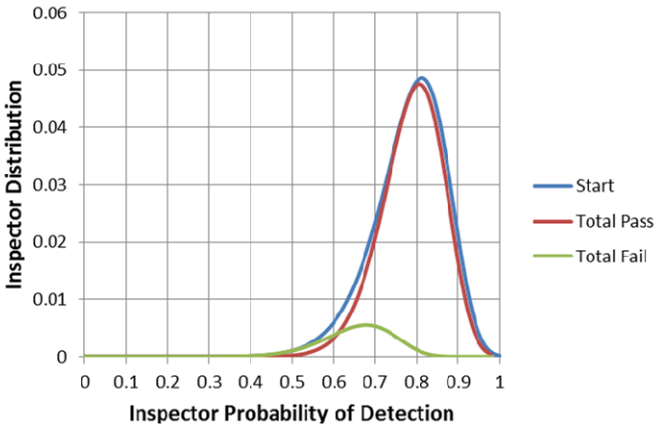


FIGURE 5: Method for Determining the Distributions of Inspectors who Pass and Fail a Hypothetical 8/10 Detection Test

This method was specifically used to answer specific questions about austenitic inspections with IGSCC. First, given a 63% pass rate for the first test and an 82% yield after three tries, what initial skill distributions would yield these results and what would the skill distribution of the inspectors in the field be after up to three rounds of testing?

The first case examined used simple Weibull and normal distributions of inspector skill levels to approximate a mixed but relatively homogenous group of inspectors. It was determined that neither the simple normal nor Weibull distribution alone could produce a 63% pass rate and an 82% yield after three tests. Ten people dropped out of the testing, with five dropping out after the first and second round, so a higher priority was given to the initial pass rate.



Initial Pass Rate
62.52%

Yield after 3 Trials
88.5%

Average POD Of Start	Median POD Of Start
78.26%	78.0%

Average POD Of Passed	Median POD Of Passed
79.8%	79.0%

Average Improvement	Median Improvement
1.5%	1.0%

FIGURE 6: Distributions for Skilled Inspectors Who Pass and Fail an 8/10 Test after Three Tries To Obtain a 63% Pass Rate

What is interesting about the case of using a Weibull distribution for the initial inspector skill levels is that the skill distribution is almost unchanged by the testing process. Some of the weakest performers would fail, but the improvement in the POD if the inspectors who would pass is only on the order of 1.5%, which is much smaller than the uncertainties associated with the initial pass rates. If one were to take the skill distribution of the inspectors who passed this test after three tries and immediately gave them one test the expected pass rate would be 66%, which is within the statistical uncertainty of the 63% initial pass rate.

The second case is one where the initial inspectors consist of two distinct groups, skilled and unskilled inspectors. This was performed to explore the anecdotal information that some licensees and inspection vendors were using the PDI testing as a de facto training program. To simulate this situation, a distribution consisting of skilled and unskilled inspectors was produced and tuned to give the 63% initial pass rate and 82% yield. It became apparent during the development of this mixed distribution that a yield of 82% prevented including a mix of more than 20% unskilled inspectors. The power curves essentially preclude inspectors with a POD of less than 50% from passing the test. A distribution using 20% of unskilled inspectors, defined as having a normal distribution centered on a 50% POD, and skilled inspectors using a Weibull distribution, is shown in Figure 7.

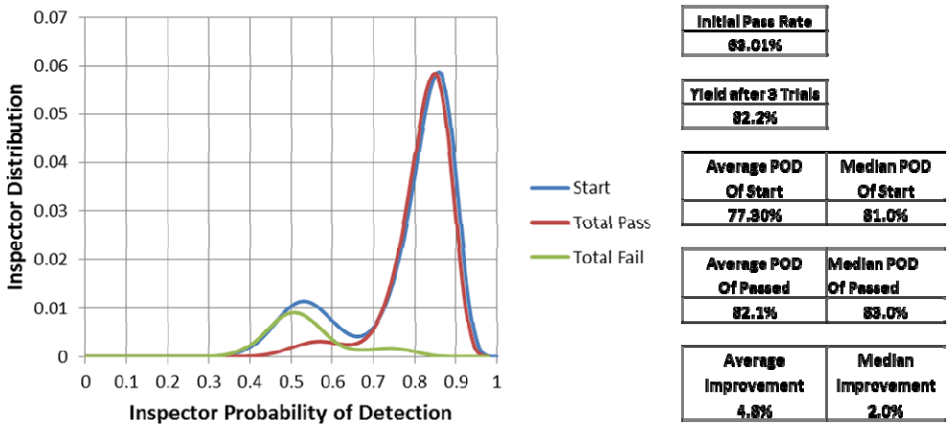


FIGURE 7: Distributions for a Mix of Skilled and Unskilled Inspectors Who Pass and Fail an 8/10 Test after Three Tries To Obtain a 63% Pass Rate

In the case of the mixed skill distributions, the improvement in initial vs. final inspector POD is 4.8%. This near 5% increase is the largest that could be obtained while maintaining the 63% initial pass rate and the 82% yield. This suggests that the influence of the initial skill distribution on the actual POD of the inspectors passing the Austenitic with IGSCC test is at most 5%. If one were to take the skill distribution of the inspectors who passed this test after three tries and immediately gave them one test the expected pass rate would be 77%, which is significantly higher than the 63% initial pass rate.

DISCUSSION

This analysis, while missing important information, such as the false call rates, was able to provide insight into inspector performance under laboratory conditions.

The estimated PODs for inspectors under laboratory conditions can be conservatively estimated at being approximately 75-85%. This POD range assumes the worst possible initial distribution of inspectors and does not take the possibility of failures due to false calls into account. The highest PODs, over 85% for encoded and non-encoded examinations was, unsurprisingly, for the easiest inspection, those for cracks in ferritic welds. One interesting aspect of the POD results was that a very challenging inspection, that for dissimilar metal welds, had a higher POD than an apparently simpler inspection, the inspection of austenitic welds. This enhanced POD is likely the result of the more complex nature of the procedures used to inspect dissimilar metal welds. Examinations of dissimilar metal welds are normally conducted with the weld crown ground or machined to allow for the ultrasonic search unit to scan directly on top of the weld, while many procedures for austenitic welds can be conducted with

the weld crown in place, which can prevent the search unit from achieving sufficient coupling while on the weld surface.

The analysis of the power curves and pass rates to determine the most likely distribution of initial inspectors was useful in evaluating the anecdotes that unskilled inspectors are being sent to PDI and sometimes getting by via luck. As the yield after three tests is around 80-100% for the various tests and the design of the test gives a ≈ 5% chance of passing for an inspector with a 50% POD, it is unlikely that the numbers of people being sent to PDI without proper preparation are getting through in any numbers. If 20% of the inspectors are unqualified, one would expect ≈3% of the inspectors in the field to be of this quality. The possibility that some number of unskilled inspectors are being sent to PDI as training cannot be precluded by the yield rate, but it does suggest that they are relatively low in number. That said, the NRC staff understands that the PDI staff could be dealing with dozens of people per year who are not properly prepared for the tests, leading to the anecdotes.

When determining the level of conservatism of the estimated PODs shown in Figure 4, the analysis of various initial inspector skill distributions appear to show that the effect of the skill distribution is only a few percent at most. The statistical uncertainties in the POD values is larger than any adjustment for the initial skill distributions.

When one looks at the requalification test pass rates for the IGSCC tests, which is slightly lower than the initial qualification rates, the hypothesis that the inspectors have a similar skill level appears to be more likely than the mixed-skill hypothesis. If a significant number of unskilled inspectors have been weeded out by the initial testing, the requalification rates would be measurably higher than the initial qualification rates, assuming the inspector skill levels remained the same between tests. The requalification rates also suggest that the inspectors are, at best, maintaining their skills and not improving them during the time period between the tests.

The concern that the IGSCC requalification rates represent a safety issue w determined to be invalid. This concern is understandable as people conducting UT, like many measurement techniques, have a “calibrate before and after testing” mentality and if a system fails to calibrate properly after a series of tests the quality of the tests is called into question. The “calibrate before and after testing” idea does not apply well to the qualification and requalification tests based on the low pass rates for the initial qualification test and the random nature of the testing. Essentially, if your system randomly fails to calibrate half the time at the beginning of a test, that it randomly fails 50% of the time at the end of the test is not only not unusual, it should be expected.

ON-GOING AND FUTURE RESEARCH

While the analysis of the PDI test rates shows that the PODs in the laboratory environment are relatively high and that the inspectors who pass the tests should have slightly better PODs than those given, this finding only applies to the PDI testing environment and not field conditions. While performance demonstration tests help to ensure that equipment, procedures and personnel are capable of reliably detecting flaws in a laboratory testing environment, notable failures have recently occurred during application in the field. Often, when failures occur, the root cause analysis cites human error as a contributing cause, thus, providing indication that field performance may be degraded as compared to PDI testing.

Research at the NRC has been initiated to identify the human factors issues that are most likely to impact personnel performance during nondestructive examinations (NDE). The research will compare laboratory-based NDE with NDE in the field in order to better understand the key differences between the human factors issues that are present in the field that are not accounted for in the laboratory environment.

This research will focus on manual conventional UT and manual phased array UT. The first step of the research will consist of a systematic comprehensive review of literature pertaining to human factors and NDE in order to determine what factors have been adequately studied and what factors require more research. Next, detailed observations of UT performance demonstration testing in the laboratory environment and field inspections will be performed to identify differences between lab and field UT in terms of existing or potential HF challenges. Finally, the human factors issues identified will be prioritized to serve as input to the development of a plan for future research and for addressing human performance challenges in UT.

REFERENCES

1. ASME, ASME Boiler and Pressure Vessel Code Section XI, *Rules for Inservice Inspection of Nuclear Power Plant Components,* July 1, 2015.
2. C. Latiolais, *PDI Qualification Statistics Update,* Industry/NRC NDE Technical Information Exchange Public Meeting, January 15, 2014, pp. 56-66. (ADAMS Accession No. ML15013A515)