Conf-921007--12

PNL-SA--21374

DE93 004328

PERFORMANCE DEMONSTRATION REQUIREMENTS FOR EDDY CURRENT STEAM GENERATOR TUBE INSPECTION

R. J. Kurtz
P. G. Heasler
C. M. Anderson

October 1992

Presented at the
20th Water Reactor Safety
Information Meeting
October 21-23, 1992
Bethesda, Maryland

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

PNL-SA-21374

PERFORMANCE DEMONSTRATION REQUIREMENTS FOR EDDY CURRENT STEAM GENERATOR TUBE INSPECTION

R. J. Kurtz, P. G. Heasler, and C. M. Anderson

Pacific Northwest Laboratory
Operated by Battelle Memorial Institute
for the U. S. Department of Energy

ABSTRACT

This paper describes the methodology used for developing performance demonstration tests for steam generator tube eddy current (ET) inspection systems. The methodology is based on statistical design principles. Implementation of a performance demonstration test based on these design principles will help to ensure that field inspection systems have a high probability of detecting and correctly sizing tube degradation. The technical basis for the ET system performance thresholds is presented. Probability of detection and flaw sizing tests are described.

1.0 INTRODUCTION

Eddy current inservice inspections (ISIs) of steam generator tubing are routinely performed as an element in the overall defense-in-depth strategy for ensuring the structural and leak-tight integrity of the reactor coolant pressure boundary. The main objectives of these inspections are to detect evidence of tube degradation so that corrective action(s) may be taken to mitigate tube damage, and to catch most or all degraded tubes that could fail by leak or burst before the next inspection. To attain these objectives a reliable ISI must be performed.

To ensure the reliability of these ISI's, performance demonstration qualification tests have been developed. A performance demonstration test should duplicate (as closely as possible) the conditions that would exist for the ET systems in the field on real steam generators. For this paper, an ET inspection system was taken as the ET personnel, equipment, and procedure in combination. During the performance demonstration test, the ET system should have no more information available than in the field. The ET system should inspect tubing containing realistic flaws and should be graded on how reliably flaws can be found and sized. The test should not be designed to evaluate intermediate steps in the inspection process, but should concentrate on the ultimate outputs, proper flaw detection and sizing.

The first step in the development process is to formulate the basic performance demonstration objectives as hypotheses tests of the form:

H_0: The ET system is unacceptable

versus

H_1: The ET system is acceptable

Statistical design calculations can then determine a proper pass/fail threshold and the most efficient grading scheme for the hypothesis test.

Although this general strategy is quite straightforward, several important issues have to be resolved before a workable test is actually constructed. These issues include:

1. How should ET system performance (or reliability) be quantified?

2. What performance "thresholds" should ET systems exceed to be considered qualified?

3. How is a test constructed to ensure with a high level of confidence that ET systems exceed the selected performance thresholds?

This list produces a framework for the construction of a performance demonstration test. Information relating to the first issue is contained in Section 2 of this report, the second issue is addressed in Section 3 and the third issue is addressed in Sections 4 through 6. It is important to note that before a statistical test for performance demonstration can be constructed, these issues must be resolved. To construct a statistical test, one must describe the test objectives in quantitative terms. Related to these issues is the matter of measuring test performance (as distinguished from ET system performance). Statisticians have standard measures for evaluating the performance of a test (called Type I and Type II errors). Consequently, after a workable test has been constructed and effort is directed on improving the test or determining the appropriate sample size of the test, it becomes important to calculate Type I and II errors for the prospective tests.

2.0 QUANTIFICATION OF ET SYSTEM PERFORMANCE

A reliable ET inspection system must perform two tasks, it must detect flaws a high percentage of the time and then accurately size them. Because of this, evaluation of detection reliability is usually separated from sizing. Detection performance is quantified by means of a probability, which is most commonly called probability of detection (POD). Sizing performance has typically been quantified in a less standard manner. Generally speaking, sizing performance is usually described by some sort of regression model which relates true flaw size to measured flaw size. Parameters, defined in terms of this regression model, are then used to measure sizing performance.

2.1 Probability of Detection

Probability of detection is defined as the probability the ET system will detect a flaw of a certain size, s, and is denoted by $POD(s)$. With the use of POD, an ET system's performance can be summarized by a curve, as illustrated in Figure 1 below. This POD curve completely describes the two types of errors an inspection system can make during the task of detection. A "significant" flaw may be missed or good material may be called flawed. If C_o is the size of a significant flaw, then $1 - POD(C_o)$ represents the chances of committing the first error, while $POD(0)$ represents the chances of committing the second error (i.e., the false call probability = $POD(0)$).

Inspections will produce binary data which can be used to estimate the POD. For example, the points displayed in Figure 1 represent binomial data obtained from inspections. Each point in this figure represents estimated POD for a collection of flaws over a small size interval. Such data can be fitted to any parameterized family of curves as the points are in Figure 1. In Figure 1, the points have been fit to a logistic curve, perhaps the most popular type of curve used to model POD.

POD curves can be used to distinguish an acceptable inspection system from an unacceptable one. It is easy to describe what the POD curve of a completely ineffective inspection system would look like, it would be a horizontal line. When the POD curve is horizontal, the inspection system has essentially the same chance of calling a flaw in good tubing as in degraded tubing. This performance is no better than guessing. In contrast, "ideal" performance may be represented by a step function. For this step function, POD = 0 for small flaws that have no safety significance and POD = 1 for flaws of "significance". An inspection system with an ideal POD curve will never make a mistake; no false calls will be made and no "significant" flaws will be missed.

In the case of steam generator tube inspection, it is important to note that "significant" flaws are not only the ones large enough to threaten tube integrity. Detection of flaws smaller than the "critical" size ("critical" in the context of this paper refers to a flaw severe enough to cause failure of the tube by leak or burst) is important because steam generator sampling strategies rely on detection of tube degradation at an early stage to aid in identifying defective tubes. For example, U. S. NRC Regulatory Guide 1.83 (U.S. NRC, 1975) and the most recent edition of the EPRI Inspection Guidelines (EPRI, 1988) have criteria which trigger additional inspection when flaws less than 40% through-wall (TW) are detected. The rationale for this strategy is that detection of flaws less than the ET plugging limit indicates the presence of a problem in the steam generator. Depending on the numbers of degraded tubes discovered, additional inspection may be required. Further, the location of all degraded tubes must be recorded and these tubes included in the sample set for the next inspection. This strategy is based on the idea that detection of tube degradation below rejectable limits alerts the owner and NRC to a potentially significant condition that needs to be followed and

aids the process of identifying tubes with rejectable flaws by examining more tubes in the steam generator.

In order to specify POD curves that are "acceptable", it is therefore natural to designate a region that is "close" to the ideal step function. POD curves that do not fall within this region would be considered unacceptable and any team that has such a POD would be considered unqualified.

Difficult compromises are involved in the determination of this region, however. The more stringent it is made, the more likely that no existing inspection system can satisfy its requirements and the more likely resources will be required to develop new detection techniques. On the other hand, the less stringent it is made, the greater the post-inspection tube failure probability. Determination of the shape of this region is therefore inherently a cost versus benefit question.

2.2 Flaw Sizing Regression Model

Sizing performance is generally evaluated through a regression model. Most commonly, flaw sizing is assumed to obey a regression relationship of the form:

$$M(t_i) = \beta_1 + \beta_2 * t_i + e_i \qquad (1)$$

where $M(t_i)$ represents the measured size of a flaw with true size t_i. According to this regression model, measured and true sizes are related to each other in a linear manner as defined by the parameters β_1 and β_2.

The error term e_i is assumed to be a normal deviate with mean 0 and constant standard deviation of σ. Although these assumptions are not often explicitly stated, they are necessary if the regression results are to give an adequate description of sizing performance. For many sizing procedures, the error distribution is skewed, and the shape of the distribution is dependent on flaw size. The error distribution for small flaws typically has a heavy right-hand tail while the situation is reversed for large flaws.

Ideal sizing performance should fit a regression model of the form:

$$M(t_i) = 0 + 1 * t_i \qquad (2)$$

In other words, ideal sizing performance exhibits $\beta_1 = 0$, $\beta_2 = 1$, and $\sigma = \text{stdev}(e) = 0$. Therefore, acceptable sizing performance should fulfill the following criteria:

1. σ should be suitably small.

2. $\beta_1 \approx 0$ and $\beta_2 \approx 1$ so there is little bias in the measurements.

3. The linear regression model should be a reasonable description of the data.

When the sizing data does not fit a linear regression model, this must be considered unacceptable performance. This may result in a non-linear relationship between the true and measured sizes or a non-normal error distribution.

A single parameter known as the mean-square-error (MSE) can be used to ensure that all of the above requirements for acceptable sizing performance are met. In fact, if one requires that mean square error is less than c^2, that is,

$$MSE(t) = E(M(t) - t)^2 < c^2 \qquad (3)$$

then the following bounds on the regression parameters must hold:

1. $\sigma < c$

2. $\beta_1 < c$ and

3. $|\beta_2 - 1| < c/t$

The MSE is therefore a very concise parameter for specifying acceptable sizing performance. In order to provide a reasonable requirement for sizing performance, one should only require a low MSE for flaws in the range from 10% to 100% TW. When flaws are smaller than 10% TW and very difficult to size, it could be an unreasonable requirement for ET systems to produce a low MSE.

3.0 ET SYSTEM PERFORMANCE THRESHOLDS

3.1 Degraded and Defective Tubes

As discussed in Section 2.0 a reliable ET inspection system must detect and accurately size safety significant flaws a high percentage of the time. In addition, an ET inspection system must also possess acceptable reliability to detect and size smaller flaws which are not safety significant but which serve to alert the owner and NRC to conditions which may require corrective actions to mitigate further tube damage. In this section we give definitions of degraded and defective tubes to provide the basis for establishing the ET system performance thresholds.

A defective tube is one which contains a flaw of such severity that the tube is unacceptable for continued service. A degraded tube is one which contains a flaw of lesser severity than a defective tube. For this work a defective

tube was defined as one with TW degradation severe enough to cause tube failure by burst under main-steam-line-break loading conditions, or by leakage under normal operating or accident loading conditions. To determine the flaw severity which would result in a tube being classified as defective, test data on tube failure pressure as a function of flaw size and geometry were utilized.

Relevant failure pressure data have been published by Alzheimer, et al. (1979) and Kurtz, et al. (1988) on mechanically and chemically flawed specimens of Inconel 600 tubing. From the data, constitutive equations were developed relating tube failure pressure to flaw size and morphology.

The burst-mode constitutive equations were used to develop a definition of an unacceptable flaw which was used in the development of the performance demonstration tests. Figure 2 shows a plot of these equations for an 0.875 x 0.050 tube with an essentially infinitely long flaw. It is evident from this plot that an 85% TW flaw represents an average depth for all flaw types that would burst under main-steam-line-break loading conditions (≈2600 psi pressure differential). If a flaw growth rate of 10% per operating cycle is assumed, then a tube with an actual flaw ≥ 75% TW flaw could fail under main-steam-line-break loading conditions by the end of the next operating period. This level of degradation was used to define an unacceptable (i.e., defective) tube condition requiring tube plugging or repair.

3.2 ET System Performance Thresholds

The purpose of the performance demonstration test is to provide a mechanism to ensure that field inspection systems (i.e. personnel, equipment and procedure) can reliably detect and size flaws in steam generator tubing. For this development effort the goal of steam generator tube ISI was to identify all defective tubes which could fail by leak or burst during reactor operation. Research work (Bowen, Heasler, and White 1989; Hanlen 1990) to develop and evaluate ISI sampling plans indicated that a 40% systematic, sequential sampling strategy was almost as effective as 100% inspection for identifying defective tubes, assuming some clustering of tube degradation. This sampling strategy relies on two key concepts to achieve this high level of effectiveness. First, a relatively large, uniformly distributed initial sample is used to provide a reasonable probability of finding isolated defective tubes, and second, detection of tube degradation of any severity triggers second-stage inspection to aid in finding defective tubes which may be in close proximity. In order for this sampling strategy to be effective good flaw detection and sizing reliability is needed even when degradation is < 75% TW.

Based on the above requirements the threshold POD curve shown in Figure 3 was selected to define unacceptable POD performance. The defining points for the threshold POD curve are listed in Table 1. This particular threshold POD curve was selected so that ET systems possessing performance characteristics at or below the threshold curve would fail the test a high percentage of the time, and ET systems with performance characteristics similar to the accept-

able curve plotted in Figure 3 and listed in Table 1 would pass the POD test a high percentage of the time. A team with "acceptable" POD performance would have a ≥ 95% probability of detecting a defective tube (flaws ≥ 75% TW) and ≥ 90% probability of detecting flaws ≥ 40% TW. It was judged from the ET reliability studies conducted on the retired-from-service steam generator (Kurtz, et. al. 1990) that the "acceptable" level of POD performance was attainable by ET systems employing state-of-the-art inspection equipment and procedures.

There are two mistakes that can be made when using test results to determine the acceptability of an ET system. The first mistake is that an ET system is called acceptable when it is really unacceptable. The probability of making this type of mistake is called the Type I error. The second mistake is that an ET system is called unacceptable when it is really acceptable. The probability of making this type of mistake is called the Type II error. The probability of correctly identifying an acceptable ET system is called the power of the test.

For this test, as with any hypothesis tesing problem, limits on the Type I and Type II errors are specified. These limits and the pass/fail threshold determine the final sample size requirements. The acceptable and unacceptable ET thresholds (given in Table 1) were used with a Type I error of 10% and Type II error of 7% to determine sample sizes for flaws 20%, 40%, and 75% TW. Sample sizes were chosen for 0% (Blanks) and 100% TW such that the overall Type I and Type II erors would be 0.001% and 30%, respectively. Monte Carlo simulations (described in Section 6), were utilized to investigate the actual Type I and Type II errors and to see the effects on the error of using alternative acceptable/unacceptable POD curves and sample sizes.

Table 1. Defining Points for Unacceptable and Acceptable POD Performance

Through-Wall Flaw Depth, %	Unacceptable POD Curve	Acceptable POD Curve
0 (Blank)	≥ 0.15	< 0.15
20	≤ 0.15	0.30
40	≤ 0.80	0.90
75	≤ 0.90	0.95
100	≤ 0.90	0.95

Similar thresholds were selected on the MSE to establish appropriate controls on flaw sizing performance. Figure 4 gives a plot of the root mean squared error (RMSE) versus PEL for teams participating in the Surry round robins

(Kurtz, et.al. 1990). The PEL quantity is the probability of a flaw being sized by an ET system in excess of the plugging limit when the tube is truly defective (i.e., with degradation $\geq$ 75% TW). Also shown in Figure 4 are results of a theoretical calculation of RMSE versus PEL. From this plot, a value of RMSE = 20 was selected to represent unacceptable sizing performance since this value of RMSE would yield a PEL of about 0.93. In other words, ET systems with RMSE $\geq$ 20 should fail the sizing test a high percentage of the time. A value of RMSE = 17 was chosen to represent acceptable sizing performance since this would produce a PEL of about 0.95. The sizing test was designed so that ET systems with RMSE $\leq$ 17 would pass the test a high percentage of the time. It should be emphasized that these values of RMSE were selected on the basis of a 40% TW plugging limit and the definition of a defective tube given above. If another plugging limit is used then different values of RMSE must be specified.

The POD and flaw sizing performance characteristics were selected so that a passing ET system would possess an overall $\geq$ 90% chance of detecting and plugging a defective tube, provided the tube was inspected. This is readily apparent since the acceptable POD performance for defective tubes is $\geq$ 95% and the acceptable sizing performance is PEL $\geq$ 95% which results in a joint probability of detecting and correctly calling a tube defective when the flaw size is $\geq$ 75% of about 90%. In addition, the POD performance of an ET system likely to pass the POD test would be about 90% for flaws $\geq$ 40% TW. Sections 5 and 6 of the report present the detailed statistical calculations that were performed to develop performance demonstration tests to meet these design objectives.

4.0 RECOMMENDED PERFORMANCE DEMONSTRATION TEST

This section describes the recommendations for the performance demonstration test, including a description of the number of tubes to be inspected, the distribution of the flaw sizes, and the methods for grading the POD and sizing performance of the ET systems. The statistical details that were used as the basis for this section are presented in Sections 5 and 6.

4.1 General Structure of Performance Demonstration Test

For a performance demonstration test, the flaw types and locations should simulate those found in operating steam generators. Specifically, the specimen set should be unknown to the personnel operating the inspection equipment in order for the test results to be indicative of ET system reliability. An effective means for simulating the flaws and conditions found in real steam generators would be to construct a tube bundle mockup. Use of a mockup would provide the needed flexibility for evaluating the reliability of new NDE techniques and procedures. To be realistic the mockup must simulate conditions which affect ET inspection reliability such as steam generator internal structure, tubesheet sludge accumulations, deposits on tube surfaces, crevice deposits, and tubing geometry variations.

The matrix of flaw tubes included in the mockup should represent those flaw types and locations associated with known tube damage mechanisms such as:

(a) Wastage/Thinning
(b) Pitting
(c) Fretting/Wear
(d) Stress Corrosion Cracking initiated on either the ID (PWSCC) or OD (ODSCC) of the tube wall surface
(e) Intergranular Attack (IGA)
(f) Erosion-Corrosion
(g) Fatigue Cracking

Where appropriate, the mockup should combine flaws with other conditions which affect flaw detection and sizing reliability. The specimen set should include, but not be limited to the following conditions:

(a) Tube expansion transitions created by rolling, hydraulic or kinetic, methods
(b) Tube bend transitions
(c) Tube support plates, egg crates, or tubesheet simulations
(d) Antivibration bars or spacers
(e) Tubesheet sludge
(f) Crevice deposits
(g) Deposits on tube surfaces

A large percentage of the flaws included in the mockup should be cracks representative of typical orientations and locations since this is the most prevalent form of tube degradation occurring at this time.

The recommended number and depth range of flaws to be included in the mockup is given in Table 2. The length of the flaws should be 0.020 inches or greater. These numbers were derived to produce approximate Type I and Type II errors of 0.10 and 0.07, respectively, for flaw detection at each individual flaw size. The statistical basis for these numbers is described in Section 5.

Assessment of POD and flaw sizing reliability requires knowledge of the true dimensions of each flaw. The processes used for producing controlled sized flaws should be validated (with respect to size) by destructive metallographic analysis of specimens. Since it is impractical to destructively measure all test specimen flaws, the group of flaws incorporated in the mockup should be nondestructively characterized prior to use for performance demonstrations. Destructive measurements should be made periodically on a percentage of the flaws to verify the accuracy of the techniques used to provide the nondestructive flaw characterization data.

Table 2. Number of Flaws for Tube Mockup

Through-Wall Flaw Depth, %	Number of Samples
0 (Blank)	100
10-30	60
31-60	90
61-90	200
100	10

4.2 POD Test Grading Methods

As described in Section 1, the performance demonstration objectives must be formulated as a hypothesis test. The hypotheses are defined in terms of two threshold values. The form of the hypotheses for this test will be:

H_0: $POD_{ET\ System}(s) \leq POD_U(s)$ for all flaw sizes $s \geq$ 20% TW

$POD_{ET\ System}(0) \geq POD_U(0)$ for blanks

versus

H_1: $POD_{ET\ System}(s) \geq POD_A(s)$ for all flaw sizes $s \geq$ 20% TW

$POD_{ET\ System}(0) \leq POD_A(0)$ for blanks

The $POD_U(s)$ identifies failing performance at each flaw size s (see Table 1) and $POD_A(s)$ identifies passing performance at each flaw size s (see Table 1). The blank specimens are considered to include a flaw of size 0 and dealt with in the same way as the other "flaws" as a way of incorporating false call information into the demonstration test. We would like ET systems with an unacceptable POD to fail the test a high percentage of the time and those with acceptable POD to pass a high percentage of the time. The specifics of how to grade the ET system and decide between hypotheses are discussed in Section 5, but are outlined in the following paragraph.

The ET system is graded by 1) estimating the POD curve, 2) calculating 80% confidence limits for the estimated curve, and 3) comparing the lower confidence limit to the threshold (unacceptable) curve, designated as $POD_U(s)$ and shown in Figure 3. A passing POD curve is one with a 80% lower confidence limit which is greater than the curve shown in Figure 3 over the interval 20% to 100% TW. A computer program has been developed to estimate the POD curve

and 80% confidence bounds from the inspection results. In addition, the false call rate must be less than or equal to 12% to pass the test.

4.3 Flaw Sizing Test Grading Methods

The sizing test is graded by calculating the RMSE of the depth measurements. The minimum number of flaws required is 170, and their sizes should be uniformly distributed over the interval 10% to 90% TW (with 10 of the total number of flaws being 100% TW). A subset of the detection test specimen set may be used for this test. The grading criteria for this and larger sample sizes is given in Table 3. The statistical background for this testing method is found in Section 6.

Table 3. Number of Flaws and Acceptance Criteria for Sizing Test

Number of Flaws	Acceptable RMSE, %
170	18.20
200	18.35
250	18.52

5.0 STATISTICAL BACKGROUND FOR THE POD TEST

This section provides the specific statistical background for the performance demonstration POD test. The objectives of the performance demonstration POD test have been expressed as hypotheses in terms of two threshold values in Section 4.2.

The specifics of how to decide between hypotheses are discussed in the subsections that follow. First a description of the test is given, then a flaw size distribution is determined. The flaw size distribution will be used as the basis of a Monte Carlo simulation to determine the power of the POD test.

5.1 General Description of Calculations

To evaluate the detection performance of an ET system, the basic strategy is to present the ET inspection system with n flaws that have sizes s_i, i=1,2,3...n. These flaws are included within a large set of specimens, such as a tube bundle mockup, which also contains blank (unflawed) specimens. By comparing the inspection results to the true state of the specimens, it is possible to summarize the detection results with a binary variable, Y_i which describes whether or not the ith flaw was detected. (i.e. $Y_i = 1$ if the ith flaw was detected and $Y_i = 0$ if it was not).

The detection test will be constructed so as to use the binary data to estimate the ET system's POD curve and the "compare it" to the POD_U shown in Figure 3 and listed in Table 1. Since the estimated curve for the inspection system cannot be exact, we will surround the curve by a confidence bound and only fail ET systems whose confidence bound is at or below the thresholds given in Table 1.

To construct this test, the most widely used procedure for analyzing binary data is employed, that of logistic regression. The term logistic regression actually refers to a general algorithm that can be used to fit curves to binary data.

A form of the logistic regression curve for this test contains three independent and unknown parameters (it can be generalized to contain any number of parameters), which give the curve enough flexibility to approximate the threshold curve defined by the values in Table 1. The mathematical form of the curve can be expressed as;

$$POD(s;\beta) = \begin{bmatrix} logit(\beta_0+\beta_1 s) \text{ for } s<40\% \\ logit(\beta_2+\beta_3 s) \text{ for } s\geq 40\% \end{bmatrix} \quad (4)$$

where $logit(z) = [1 + exp(-z)]^{-1}$ and the parameters are constrained so that the curve is continuous at 40% TW flaw size. In other words, the above formulation produces a "linear" logistic curve with a possible "kink" in the curve at 40% TW flaw depth.

5.2 Approximate Flaw Size Distribution Calculations

In order to determine the approximate number of specimens needed for the performance demonstration tests, we examined binomial tests at fixed flaw sizes (20%, 40%, and 75% TW). It is recognized that in an actual performance demonstration test, ET systems would be exposed to a continuum of flaw sizes. Actual flaw sizes would range from a low of 10% up to TW. Flaw lengths would also be variable. However, the sample size determination for the binomial tests should behave approximately like the logistic test since the logistic test also considers binary data as the response, but on a flaw by flaw basis.

A sample size at each TW depth listed in Table 1 must be determined. To determine the sample size n that satisfies a particular set of Type I and II requirements, one must solve the following two binomial equations:

$$Type\ I \geq \sum_{i=1}^{n} \binom{N}{i} (P_U)^i (1-P_U)^{N-i} \quad (5)$$

and

$$Type\ II \geq 1 - \sum_{i=1}^{n} \binom{N}{i} (P_A)^i (1-P_A)^{N-i} \qquad (6)$$

These equations were solved iteratively and the results for TW depths 20%, 40%, and 75% are presented in Table 4 for Type I = 0.10 and Type II = 0.07. The discrete points from the POD_U curve and the POD_A curve given in Table 1 were used in these calculations.

Table 4. Sample Sizes for Type I = 0.10 and Type II = 0.07

Through-Wall Flaw Depth, %	Number of Samples
0 (Blank)	100
20	57
40	94
75	203
100	10

The number of blanks to be examined was chosen to be 100 to represent approximately 1/3 of the total number of flaws of size 20 to 75 % TW. In general, it is desirable to have 1/3 to 1/2 of the total number of flaws be blanks. Using Equations 9 and 10 to calculate the number of specimens with flaws of 100% TW would give approximately 400. Even though 400 specimens at 100% TW would not be feasible for this test, these flaws are represented in the test with 10 specimens (10 was arbitrarily chosen).

5.3 Evaluation of the True Errors of the POD Test

Monte Carlo simulation techniques were utilized in order to evaluate the true errors of the POD test derived in Section 5.2. A fixed sample of flaw sizes was produced according to the sample sizes determined in Section 5.2 for Type I = 0.10 and Type II = 0.07. Specifically, there were 100 blanks, 60 flaws randomly distributed between flaw sizes 10% and 30% TW, 90 flaws randomly distributed between flaw sizes 31% and 60% TW, 200 flaws randomly distributed between flaw sizes 61% and 90% TW, and 10 flaws 100% TW.

There were five unacceptable POD curves and four acceptable POD curves used in the simulations, each representing the true POD of an ET system that might be participating in the performance demonstration test. The "base case" POD for

a unacceptable ET system is the pass/fail threshold, POD_U. The "base case" POD for a acceptable ET system is POD_A, for all non-zero flaw sizes and a 5% false call rate. These are listed in Table 5.

There were four other unacceptable ET systems considered. Unacceptable System (US) #2 represents a system that has a better POD than the base case for all flaw sizes, but has an unacceptable false call rate. US #3 represents a system that handles false calls and large flaw sizes well, but has a difficult time detecting the smaller flaw sizes; i.e., performs like the base case for 20% and 40% TW. US #4 represents a system that handles false calls and small flaw sizes well, but has a difficult time detecting the larger flaw sizes; i.e., performs like the base case for 75% and 100% TW. US #5 represents a system that performs like the base case for all flaw sizes except one (20% TW for this case) where it does well.

The probability of detection for the acceptable systems were chosen to represent systems that we would expect to pass during the demonstration tests. These simulations will also help identify any biases that are introduced to the test through the estimation procedure. Acceptable System (AS) #2 represents a system whose POD is above POD_U but slightly worse than POD_A except for flaws of size 100% TW. AS #3 has a constant ability to detect flaws of size 40% TW and greater. It was of interest to see if the modeling techniques would provide confidence bounds that would fail this team a high percentage of the time. AS #4 represented a system whose POD for flaws of size 20% TW was much greater that the pass/fail threshold. This was another test of the modeling techniques.

For each of the true PODs the following steps were taken.

1) The regression parameters for the true POD were calculated based on five knot points at 0, 0.2, 0.4, 0.75, and 1.0.

2) The true POD was calculated for the flaw size distribution and then compared to a random uniform. If the POD value was larger, the flaw was designated as found. If the POD value was smaller than the random uniform value then the flaw was not found.

3) A "new" POD curve was calculated based on three knot points at 0, 0.4, and 1.0 TW with the simulated test data and then compared at five points to the POD_U.

4) The POD curve failed if it failed at every knot point. The simulations were run 1000 times and the percentage of times the simulated POD curve did not fail was tabulated.

A compilation of the Monte Carlo results using 70%, 80%, and 90% confidence bounds on the simulated POD curves is given in Tables 6, 7, and 8, respectively.

Table 5. POD Curves Used in Monte Carlo Simulations

Flaw Size, %TW	POD_U	Unacceptable					Acceptable			
		Base	2	3	4	5	Base	2	3	4
0% (blank)	15	15	20	5	5	15	5	5	5	5
20%	15	15	30	15	30	40	30	25	30	65
40%	80	80	90	80	85	80	90	85	95	90
75%	90	90	97	97	90	90	95	95	95	97
100%	90	90	97	99	90	90	95	99	95	97

Table 6. Monte Carlo Results Using 70% Confidence Bounds on Simulated POD, Percentage of Passing, Number of Simulations = 1000

Teams	Through-Wall Depth (%)					Overall
	0	20	40	75	100	
Base Case	0.610	1	0.002	0.060	0.778	0
US #2	0.066	1	0.370	0.998	1	0.034
US #3	1	0.860	0.018	1	1	0.016
US #4	1	0.992	0.534	0.074	0.282	0.042
US #5	0.236	1	0.116	0.052	0.452	0.002
Base Case	1	1	0.828	0.936	0.924	0.744
AS #2	1	0.974	0.294	0.990	1	0.286
AS #3	1	1	0.996	0.986	0.842	0.838
AS #4	0.984	1	1	1	0.938	0.922

Table 7. Monte Carlo Results Using 80% Confidence Bounds on Simulated POD, Percentage of Passing, Number of Simulations = 1000

Teams	Through-Wall Depth (%)					Overall
	0	20	40	75	100	
Base Case	0.508	1	0.002	0.044	0.678	0
US #2	0.044	1	0.272	0.996	1	0.022
US #3	1	0.816	0.008	0.998	1	0.006
US #4	0.998	0.992	0.416	0.042	0.198	0.014
US #5	0.148	1	0.066	0.030	0.356	0
Base Case	1	1	0.764	0.894	0.870	0.628
AS #2	1	0.966	0.198	0.966	1	0.182
AS #3	1	1	0.988	0.970	0.744	0.730
AS #4	0.972	1	0.998	1	0.894	0.866

Table 8. Monte Carlo Results Using 90% Confidence Bounds on Simulated POD, Percentage of Passing, Number of Simulations = 1000

Teams	Through-Wall Depth (%)					Overall
	0	20	40	75	100	
Base Case	0.344	1	0	0.016	0.528	0
US #2	0.024	1	0.150	0.982	0.992	0.006
US #3	0.996	0.710	0.006	0.996	1	0.004
US #4	0.992	0.986	0.270	0.018	0.108	0.002
US #5	0.080	1	0.038	0.006	0.216	0
Base Case	1	0.998	0.626	0.780	0.754	0.392
AS #2	0.998	0.954	0.104	0.928	0.982	0.094
AS #3	0.996	1	0.972	0.908	0.616	0.586
AS #4	0.928	1	0.996	0.996	0.754	0.686

Based on the information in Tables 6, 7, and 8, an 80% confidence bound with a sample size of 100 blanks and 360 non-zero flaws provides at most a 2% chance of passing a unacceptable system (Type I error) and approximately a 63% chance of passing the base case acceptable system. These results do not justify reduced sample sizes since that would in turn reduce the probability of passing an acceptable team.

5.4 Alternative POD Test Strategies

A 2% Type I error may be considered to be conservative for this demonstration test. It is possible that anything less than or equal to a 5% Type I error rate would be acceptable. The tables in Section 5.3 show that with a 70% confidence bound, the Type I error is still acceptable at approximately 4% while the probability of passing the base case acceptable team is approximately 74%. A reduction of sample size for this set of conditions may be acceptable since the Type I error rate is expected to remain constant while the probability of passing a acceptable team could decrease to something around 60% and still be acceptable. The results after sample size reduction with 70% confidence bounds are given in Table 9. Note that AS #2 (POD performance is slightly less that POD_A) has a very poor chance of passing under this scenario. However, this scenario could be an acceptable alternative to the POD test recommended in Section 4.

6.0 SIZING TEST

Test design calculations were also performed to determine the number of test specimens required to demonstrate that ET system capability for flaw sizing would result in about a 95% chance of calling a tube defective when the true flaw size was $\geq$ 75% TW and the plugging limit was 40% TW. It was assumed that the objective of the test is to distinguish between two hypotheses of the form:

H_0: $MSE > (\sigma_u)^2$

versus

H_1: $MSE < (\sigma_a)^2$

where sizing performance is measured by the MSE (see Section 2.2). The MSE is defined by the formula;

Table 9. Monte Carlo Results Using 70% Confidence Bounds on Simulated POD, Sample Size is 75% of 360 Sample Test, Percentage of Passing, Number of Simulations = 1000

Teams	Through-Wall Depth (%)					Overall
	0	20	40	75	100	
Base Case	0.578	1	0	0.074	0.712	0
US #2	0.080	1	0.258	0.992	1	0.030
US #3	0.998	0.800	0.032	0.998	1.000	0.024
US #4	0.992	0.992	0.380	0.084	0.284	0.024
US #5	0.228	1	0.132	0.066	0.414	0.004
Base Case	1	0.996	0.734	0.896	0.864	0.580
AS #2	0.998	0.962	0.194	0.986	1	0.188
AS #3	0.998	0.998	0.978	0.956	0.792	0.766
AS #4	0.962	1	0.990	0.998	0.898	0.856

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \{M(t_i) - t_i\}^2 \qquad (7)$$

where M_i is the measured flaw size and t_i is the true flaw size. The MSE for a particular flaw size is related to the bias and standard deviation according to the formula;

$$MSE = \sigma^2 + B^2 \qquad (8)$$

The test is defined in terms of two threshold values σ_u, which identifies unacceptable performance and σ_a which identifies acceptable performance. The relationship of these thresholds to PEL is discussed in Section 3.2 and plotted in Figure 4. The test was designed so that ET systems with unacceptable MSE would fail the test a high percentage of the time and those with acceptable MSE would pass the test a high percentage of the time.

The test is conducted by having the ET system size n flaws. It should be noted that there would not be two separate performance demonstration tests, one for detection and one for sizing. ET systems would inspect one tube

mock-up and be required to report both detections and size those flaws detected. The flaw sizes are randomly distributed within the sizing region of interest. The ET system passes if the MSE is less than the critical value, *c* and fails otherwise. The objective of these calculations is to determine reasonable values for *n* and *c*. Table 10 presents sample size requirements for Type I = 0.05 and Type II = 0.10.

Table 10. Sample Size Requirements for Flaw Sizing Test, Type I = 0.05, Type II = 0.10

σ_1	Type I Pr(Pass)	σ_2	1 - Type II Pr(Pass)	Number of Samples	Pass Criteria
20	0.05	17	0.90	170	18.20
20	0.05	17	0.90	200	18.35
20	0.05	17	0.90	250	18.52

It is evident that the number of samples needed to conduct an adequate performance demonstration test for flaw sizing is considerably smaller than for the POD test.

7.0 SUMMARY AND CONCLUSIONS

Statistically based performance demonstration qualification requirements have been developed to ensure that field ET inspection systems can reliably detect and size all of the known forms of tube damage that occur in operating steam generators. For this work the goal of steam generator tube ISI was to identify most or all defective tubes which could fail by leak or burst during reactor operation. An extensive data base on the failure pressure of degraded steam generator tubes as a function of flaw type and size was utilized to define a defective tube as one with degradation ≥75% TW. Information from a study on the reliability of ET systems to detect and size service-induced tube degradation, coupled with results from an effort to develop and evaluate sampling plans for ISI was used to select thresholds on POD performance, flaw sizing accuracy, and the false call rate. Thresholds were selected such that a team likely to pass the test would have a 90% composite probability of detecting and plugging a defective tube, provided the tube was inspected. Thresholds were also established for degraded but not defective tubes because current and proposed ISI sampling plans rely on detection of low levels of tube degradation to trigger additional inspection, and to alert the owner and NRC to conditions which may require corrective actions to mitigate further tube damage.

The POD, flaw sizing, and false call rate thresholds were used in statistical test design calculations to determine the appropriate number and size distri-

bution of flawed steam generator tube samples that would be needed in a steam generator tube bundle mockup to ensure reliable ET inspection system performance. Binomial calculations and Monte Carlo simulations were performed for mockups containing different numbers and minor variations of a particular distribution of flawed tube samples to determine the probability of an acceptable ET system failing the test and for an unacceptable ET system passing the test. For the POD test a mockup consisting of 360 flawed tube samples would be needed to meet the performance goals selected. A computer program has been developed for grading the POD test. For the flaw sizing test only about 170 flawed tube samples are needed to establish acceptable sizing performance.

8.0 REFERENCES

Alzheimer, J. M., R. A. Clark, C. J. Morris, and M. Vagins. 1979. Steam Generator Tube Integrity Proram Phase I Report. NUREG/CR-0718, PNL-2937, Pacific Northwest Laboratory, Richland, Washington.

Bowen, W. M., P. G. Heasler, and R. B. White. 1989. Evaluation of Sampling Plans for Inservice Inspection of Steam Generator Tubes: Part I. NUREG/CR-5161, PNL-6462, Pacific Northwest Laboratory, Richland, Washington.

Electric Power Research Institute. 1988. PWR Steam Generator Inspection Guidelines: Revision 2, 1988, EPRI NP-6201, Prepared by the Electric Power Research Institute, Palo Alto, California.

Hanlen, R.C. 1990. Evaluation of Sampling Schemes for In-Service Inspection of Steam Generator Tubing, EPRI NP-6774, Prepared by Battelle, Pacific Northwest Laboratories, Richland, Washington.

Kurtz, R. J., R. A. Clark, E. R. Bradley, W. M. Bowen, P. G. Doctor, F. A. Simonen, and R. H. Ferris. 1990. Steam Generator Tube Integrity Program/Steam Generator Group Project - Final Summary Report. NUREG/CR-5117, PNL-6446, Pacific Northwest Laboratory, Richland, Washington.

Kurtz, R. J., J. M. Alzheimer, R. L. Bickford, R. A. Clark, C. J. Morris, F. A. Simonen, and K. R. Wheeler. 1988. Steam Generator Tube Integrity Program Phase II Final Report. NUREG/CR-2336, PNL-4008, Pacific Northwest Laboratory, Richland, Washington.

U.S. Nuclear Regulatory Commission. 1975. Inservice Inspection of Pressurized Water Reactor Steam Generator Tubes. Regualtory guide 1.83, Rev. 1., washington D.C.

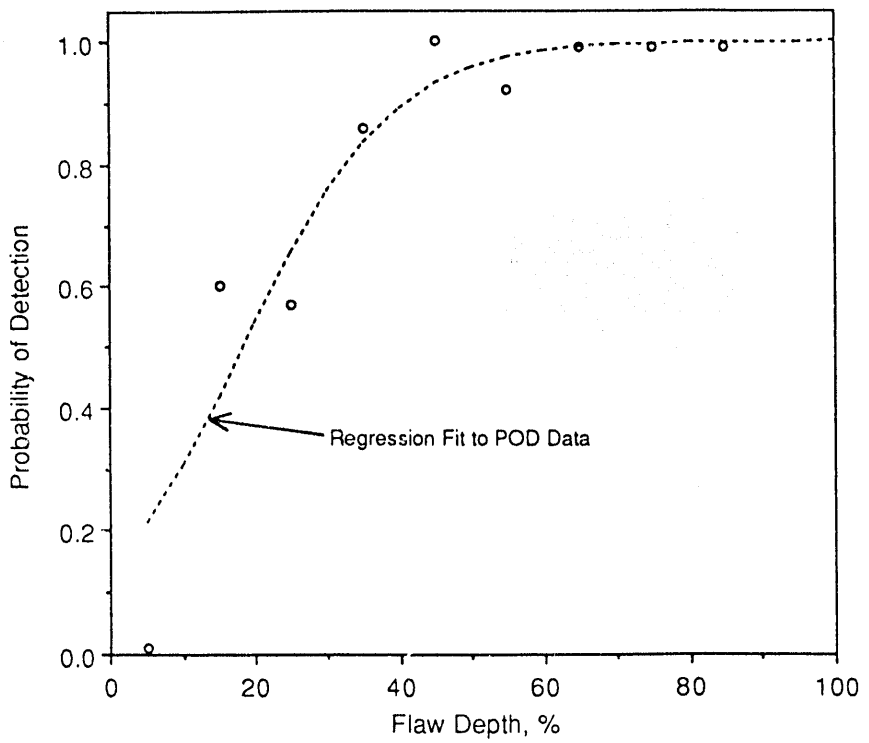

Fig. 1 Example POD Curve

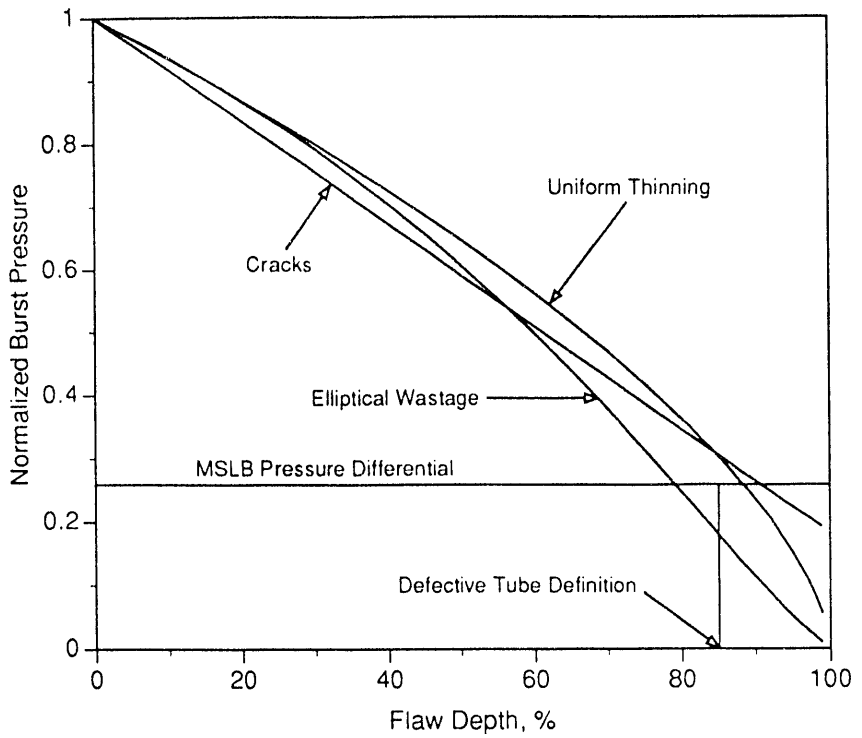

Fig. 2 Plot of Burst Mode Failure Pressure Equations Versus Flaw Depth for 0.875 in. OD x 0.050 in. [illegible] = 0.875 in

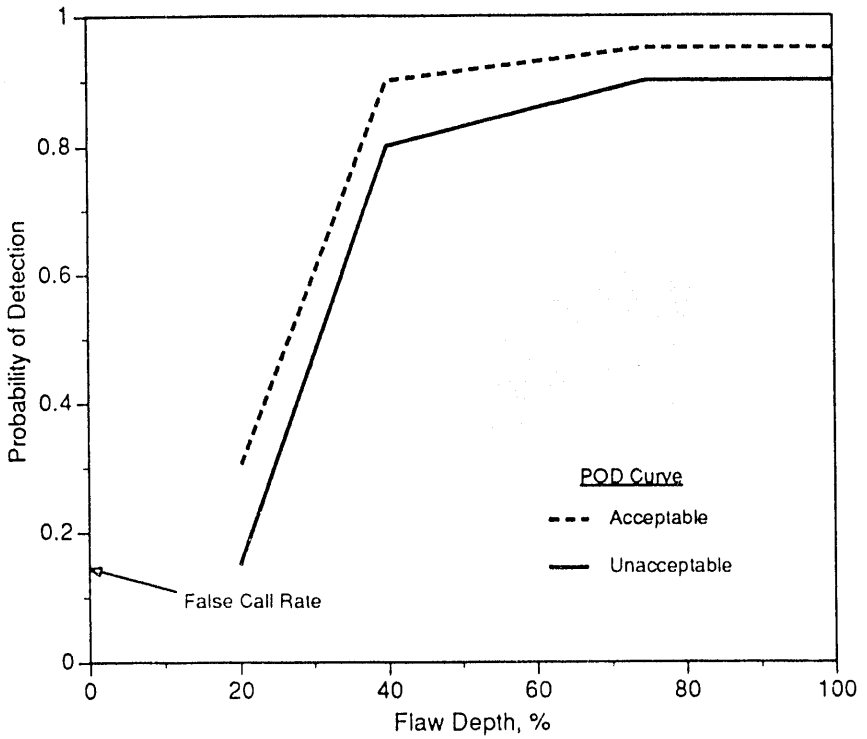

Fig. 3 Base Case Threshold POD Curves

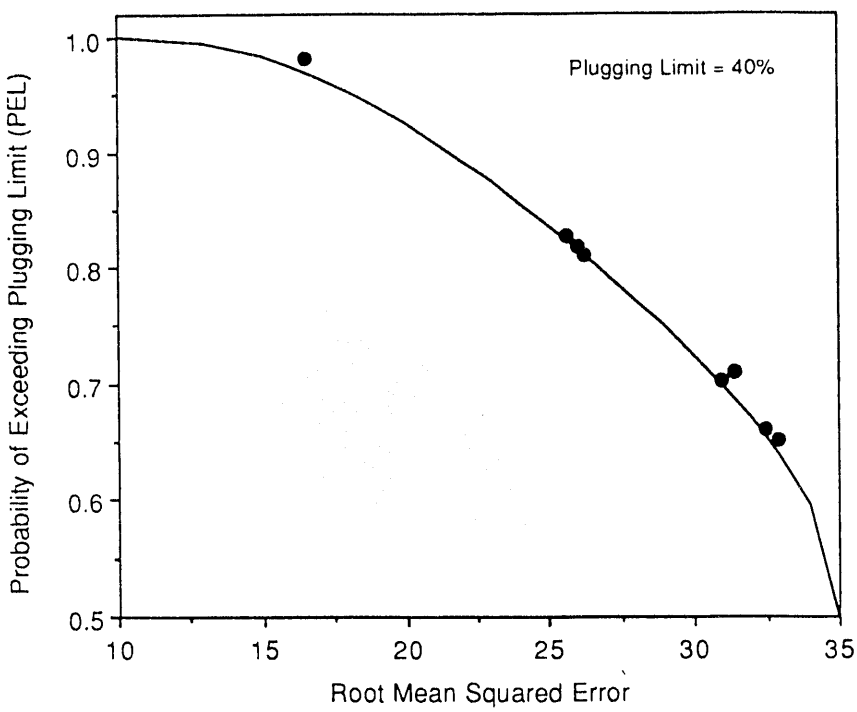

Fig. 4 Curve is Theoretical PEL Versus RMSE for 40% Plugging Limit and Defective Tube Defined as ≥ 75% TW Degradation. Data Points Represent Actual Performance of Teams Participating in Surry ET Reliability Study

END

DATE FILMED

3/2/93