EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Update on the Tools for Integrity Assessment Project

Update on the Tools for Integrity Assessment Project

1014756

Final Report, February 2007

EPRI Project Manager
H. Cothron

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

ORGANIZATION(S) THAT PREPARED THIS DOCUMENT

Westinghouse Electric Company, LLC

Aptech Engineering Service, Inc.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

CITATIONS

This report was prepared by

Westinghouse Electric Company, LLC
Nuclear Service Division
P.O. Box 158
Madison, PA 15663

Principal Investigator
T. Pitterle
R. Keating

Aptech Engineering Service, Inc.
601 W. California Ave
Sunnyvale, CA 94086-4831

Principal Investigator
B. Woodman
S. Brown

This report describes research sponsored by the Electric Power Research Institute (EPRI).

The report is a corporate document that should be cited in the literature in the following manner:

Update on the Tools for Integrity Assessment Project. EPRI, Palo Alto, CA: 2007. 1014756.

PRODUCT DESCRIPTION

This report is a revision to the data sufficiency requirements for conducting multiple nondestructive evaluation (NDE) analyst performance tests to obtain probability of detection (POD) and NDE sizing uncertainty correlations. This report was prepared to reassess the number of NDE analysis teams required for performance testing following the completion of initial performance testing for axial outside diameter stress corrosion cracking (ODSCC).

Results and Findings

The recommendation in the original EPRI report (1012985) was to use 10 analysis teams for detection and sizing testing; this report concludes that 5 teams are sufficient. Using the guidelines provided in the original EPRI report, 10 teams were used for the first performance test. However, organizing 10 teams for testing was difficult and very costly, prompting the reassessment of the number of analysis teams required. The original evaluations of the number of detection teams are retained in this report.

Challenges and Objectives

The data sufficiency requirements developed in this report for conducting performance tests include the following:

- Number of specimens for detection testing (Section 2)
- Number of NDE analysis (resolution) teams for detection testing (Section 2)
- Limitations on the number of undetected flawed specimens for detection testing (Section 3)
- Number of non-flawed specimens for detection testing for false call assessments (Section 4)
- Number and depth distribution of specimens for NDE sizing testing (Section 5)
- Number of NDE analysts for sizing testing (Section 6)
- Reassessment of number of teams required for detection testing (Section 7)
- Reassessment of number of teams required for NDE sizing testing (Section 8)
- Summary of data sufficiency requirements (Section 9)

These requirements are developed based on statistical considerations to reduce uncertainties associated with the number of specimens and analysts for POD and sizing correlations.

Applications, Value, and Use

This report will be used in future performance demonstrations under the Tools for Integrity Assessment Project.

EPRI Perspective

EPRI is uniquely qualified to carry out the remainder of this project. Using this report and the remainder of the guidelines in the EPRI report 1012985, all examination technique specification sheets will be updated.

Approach

Recommendations in this report are based on statistical analyses of the influence of uncertainties (at 95% confidence for a varying number of teams) on the depth for a POD of 0.90. The results of a completed 10-team performance test were then used to assess the sensitivity of the PODs to a reduced number of analysis teams. The influence of uncertainties on the POD—both at a given depth and on the depth at a given POD—is also addressed.

Keywords

Tools
ODSCC
Uncertainties
POD
Performance demonstration

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1
INTRODUCTION

This report develops data sufficiency requirements for conducting multiple NDE analyst performance testing to obtain probability of detection (POD) and NDE sizing uncertainty correlations. Revision 1 to this report was prepared to reassess the number of NDE analysis teams required for performance testing following completion of initial performance testing for axial ODSCC. The prior recommendation of 10 analysis teams for detection and sizing testing was implemented for the completed performance tests. This experience identified difficulties in organizing 10 teams for testing as well as high costs for this extensive testing, which are the primary reasons for the reassessment of the number of required analysis teams in this report revision. The original evaluations for the number of detection teams in Section 2 and the number of sizing teams in Section 6 are retained in this report revision. The reassessments for the number of detection and sizing teams are included as new Sections 7 and 8. This effort was performed under the EPRI Tools for Tube Integrity Assessment Program.

The data sufficiency requirements developed in this report for conducting performance tests include the following:

- Number of specimens for detection testing (Section 2)
- Number of NDE analysis (resolution) teams for detection testing (Section 2)
- Limitations on number of undetected flawed specimens for detection testing (Section 3)
- Number of non-flawed NDD specimens for detection testing for false call assessments (Section 4)
- Number and depth distribution of specimens for NDE sizing testing (Section 5)
- Number of NDE analysts for sizing testing (Section 6)
- Reassessment of number of teams required for detection testing (Section 7)
- Reassessment of number of teams required for NDE sizing testing (Section 8)
- Summary of data sufficiency requirements (Section 9)

The above requirements are developed based on statistical considerations to reduce uncertainties associated with the number of specimens and analysts for POD and sizing correlations.

2
NUMBER OF SPECIMENS AND NDE ANALYSIS TEAMS FOR DETECTION TESTING

This section provides an assessment of the impact of sample size on the intrinsic accuracy of POD functions inferred from ETSS and SSPD data sets using the logistic regression process. The work undertaken had a twofold purpose. The first was to obtain confidence bands on the inferred POD function given a specific data set. The second was to develop a quantitative basis for determination of the appropriate data set size and the recommended number of NDE resolution teams for performance testing.

In addition to addressing the general question of the effects of sample size on accuracy, the work was focused on providing POD uncertainty related input for performance of Operational Assessments (OA's), both probabilistic and deterministic in nature. In the case of the deterministic OA, the role of POD uncertainty relates to the evaluation of a hypothetical worst flaw left in service subsequent to the examination. One accepted procedure is to assume that this flaw has a depth corresponding to a 95% POD for the applicable inspection process. The uncertainty of this depth was assessed as part of this program. For the probabilistic OA, a description of the uncertainty in the overall POD function is required. This was achieved by addressing the issue of variation in the parameters defining the POD function. This is equivalent to defining the joint bivariate distribution function for the two required parameters.

Section 2.1 describes the fundamentals of standard logistic regression and the extension of this process to data sets in which the inspection of a given flaw is replicated by additional resolution teams. The beneficial effect of even a small number of replications on the POD uncertainty is shown. Section 2.2 details the process by which the POD uncertainties are assessed. The use of the Bootstrap Monte-Carlo technique to evaluate the relevant POD uncertainties is discussed together with the results for a typical case. Section 2.3 describes the results of parametric studies evaluating the uncertainties associated with a spectrum of base POD functions shown in Figure 2-1. The effects of number of flaws and number of resolution teams are evaluated providing a basis for recommended sample size requirements.

2.1 Logistic Regression

The regression process is a basic part of empirical model development. In the case of the most commonly used version, a straight line is fitted to describe a dependent variable in terms of one or more independent variables. The best set of parameters consisting of an intercept and a slope is inferred from a data set by minimizing the sum of squares of errors between observations and predictions for the data set.

Logistic regression is simply a special case of this process in which the observations of the independent variable are dichotomous (hit/miss). The parameters of the logistic function are obtained from the data set by optimization of a likelihood function rather than least-squares. The objective remains the same as for any regression process, the estimation of a best-fit model.

2.2 Estimation of Parameters

The logistic model for POD consists of 2 parameters and is given by:

$$POD = [1 + EXP(A + B \times X)]^{-1}$$

where A and B are parameters and X is the independent variable (or LOG(X) for log-logistic function).

For the basic process consisting of a single observation per data point, the likelihood function is:

$$Log(L) = \sum_i y_i \times Log(POD_i) + (1 - y_i) \times Log(1 - POD_i)$$

where Y is i^{th} data point and POD is estimated for i^{th} data point using parameters A,B.

Since the observations (y) can only take on the values of 0 or 1, only one of the two terms will be present in the summation for each data point. An example of logistic regression with no replication is shown in Figure 2-2.

2.3 Extended Logistic Regression

The logistic regression process can be extended to include repeated observations for each data point by independent resolution teams. This is accomplished by modifying the likelihood function to include more than one trial. The observation value becomes the sum of detections over all teams and the likelihood is modified accordingly:

$$Log(L) = \sum_i Binomial(N_i, POD_i, M)$$

where: N = Total Detections for all Teams

POD is estimated for i^{th} data point on per team basis

Binomial( ) = Binomial probability of N given M trials each with probability POD

M = Number of teams

An example of multi-team logistic regression is illustrated in Figure 2-3 for an analogous case to that in Figure 2-1. In this case, the use of three resolution teams allows the dependent variable to take on additional discrete values (2 & 3). This has the effect of constraining the reasonable set of parameters by reduction in the variance of the parameter estimates compared to the single

team case. This extended formulation will allow computation of a composite POD function for several resolution teams with corresponding improvement in accuracy.

2.4 Bootstrap Monte-Carlo

The regression process results in estimates of a best set of parameter estimates. These, however, have associated uncertainties which reflect limitations of the data set in representing all possible outcomes of the underlying experiment. Small samples are subject to large sampling errors reflected in higher uncertainties in the resulting parameter estimates. This section details the numerical process used to assess the parameter uncertainties associated with the logistic regression process. The logistic regression parameter uncertainties are extended to provide confidence bounds on the POD function as well as attributes important to both deterministic and probabilistic Operational Assessments.

2.4.1 Bootstrap Approach

The Bootstrap Monte-Carlo approach is a robust analog method for evaluating sampling error in the model building process. The process as shown in Figure 2-4 consists of using an assumed 'true' POD function to generate multiple synthetic data sets by a sampling process. This can be done for any desired sample size with or without multiple data point replication. For each synthetic data set, the logistic regression process is used to obtain a set of optimum parameters for the POD function. Using large numbers of such synthetic data sets provides a detailed picture of the variability of the parameter sets relative to the 'true' parameter set. The synthetic data sets were obtained by first selecting a random sample of flaw depths on the interval 0-100%TW. The true POD function was used to assign a POD value to each data point. Detection versus non-detection for each synthetic data point was by random choice. Approximately 5000 data sets were used in each simulation corresponding to 5000 estimated POD functions. From these, statistics of interest were examined.

2.4.2 Typical Case

The sample case was developed for POD function 1 (Figure 2-1) and consisted of 30 flaws and one resolution team. Figure 2-5 shows the basic simulation output in terms of the marginal distributions and dependence of the two logistic parameters. Each point in Figure 2-5 represents one realization or statistically valid POD function obtainable by sampling from the true POD. Figure 2-6 shows the result of changing the parameter definitions for the logistic function to a location and scale parameter. The resulting function is mathematically equivalent. The advantage over the initial parameter definitions is the near-normality and independence of the resulting parameter set. This facilitates the use of the simulation results for the development of a sampling function to be used in probabilistic OA's.

Figure 2-7 shows the distribution of the 90^{th} percentile POD in terms of flaw depth for the example case. This can be used to provide confidence estimates for limiting undetected flaw depth in deterministic OA's. The lower one-sided 95% confidence band on POD is shown for this case in Figure 2-8.

2.5 Parametric Study on Sample Size and Number of Resolution Teams

Parametric studies in sample size and number of resolution teams were performed using the methods discussed above to determine requirements to minimize the effect of sampling error on the uncertainty associated with the inference of POD from data. Sample sizes in terms of number of flaws varied from 15 to 100. The number of resolution teams studied ranged from 1 to 40. A varied selection of base POD functions (Figure 2-1) was evaluated to determine sensitivity in sample requirements. A full factorial study was performed for the first POD. More limited and focused studies were performed for the remaining two base POD functions.

2.5.1 POD 1

The first POD function studied was a logistic functional representation of a bobbin inspection process. This depth based function approaches a POD of unity fairly rapidly at a flaw peak depth of approximately 80%TW.

The effect of sample size in terms of number of flaws in the data set is shown in Figures 2-9 and 2-10. Figure 2-9 shows the lower 95% confidence bands for various sample sizes. Figure 2-10 shows the lower 95% confidence POD for a flaw depth of 80%TW as a function of number of flaws (1 resolution team). As can be seen from the figure, a significant improvement in the estimate results from increasing the number of flaws from 15 to 40. By comparison, the improvement by further increasing the sample size to 100 shows a decreasing return.

The effect of increasing the number of resolution teams is shown in Figure 2-11. In this case the figure of merit for the study was the upper 95% confidence estimate of the flaw depth at which the POD was 90%. This is an appropriate figure of merit for the deterministic OA application. As can be seen from the figure, the most significant improvement is predicted prior to about 10 to 15 resolution teams.

2.5.2 POD 2

The second POD function studied was a log-logistic functional representation of a +Point inspection process. This depth based function approaches a POD of unity more slowly at a flaw peak depth of approximately 50%TW.

The effect of sample size in terms of number of flaws in the data set is shown in Figure 2-12. The figure shows the lower 95% confidence POD for a flaw depth of 80%TW as a function of number of flaws (1 resolution team). As can be seen from the figure, a significant improvement in the estimate results from increasing the number of flaws from 15 to 30. The behavior is consistent with that observed for POD 1.

The effect of increasing the number of resolution teams is shown in Figure 2-13. The figure of merit for the study was the upper 95% confidence estimate of the flaw depth at which the POD was 90%. This is an appropriate figure of merit for the deterministic OA application. The behavior is consistent with that observed for POD 1.

2.5.3 POD 3

The final POD function studied was a Log-Logistic functional representation chosen for its adversity. This depth based function does not approach a POD of unity. The POD approaches a POD of 93% rather slowly. A POD function of this nature is expected to be particularly problematical in terms of tolerance of sampling error.

The effect of sample size in terms of number of flaws in the data set is shown in Figure 2-14. The figure shows the lower 95% confidence POD for a flaw depth of 80%TW as a function of number of flaws (1 resolution team). The behavior is consistent with that for POD 1 and POD 2 in that the most significant improvement in the estimate results from increasing the number of flaws from 15 to 40. The improvement by further increasing the sample size to 100, shows decreasing return. However, the lower detection performance may require increased sample size relative to the other POD functions to accommodate the sampling error.

2.6 Conclusions from Parametric Study

The beneficial effect of increasing the number of flaws in the logistic regression process has been demonstrated. The increase in benefit appears to saturate beyond 30 to 40 flaws for a single resolution group analysis. The benefit is consistent for a wide range of POD functions.

The effect of multiple resolution teams on inference of POD is significant for small numbers of teams. The beneficial effect appears to saturate beyond 10 to 15 resolution teams. POD correlations are to be based on the results of multiple analyst testing simulating field inspections for which the call is based on the results of the resolution process (analysis team). The number of analysis teams required for an ETSS dataset to support POD development should be sufficient to control the sensitivity of the POD uncertainty to the number of specimens. There is an interaction on POD uncertainties between the number of analysis teams and the number of specimens. If the ETSS approaches the minimum number of flaws, it is recommended that the number of analyst teams be increased beyond the minimums described herein to help offset the increased uncertainty from the small number of flaws. The number of analysis teams for detection testing is recommended to be ≥ 10 for a generic ETSS program. If the performance testing is being conducted for a site specific POD based on testing of the analysts used in the inspection, the number of analysis teams should be a minimum of 5.

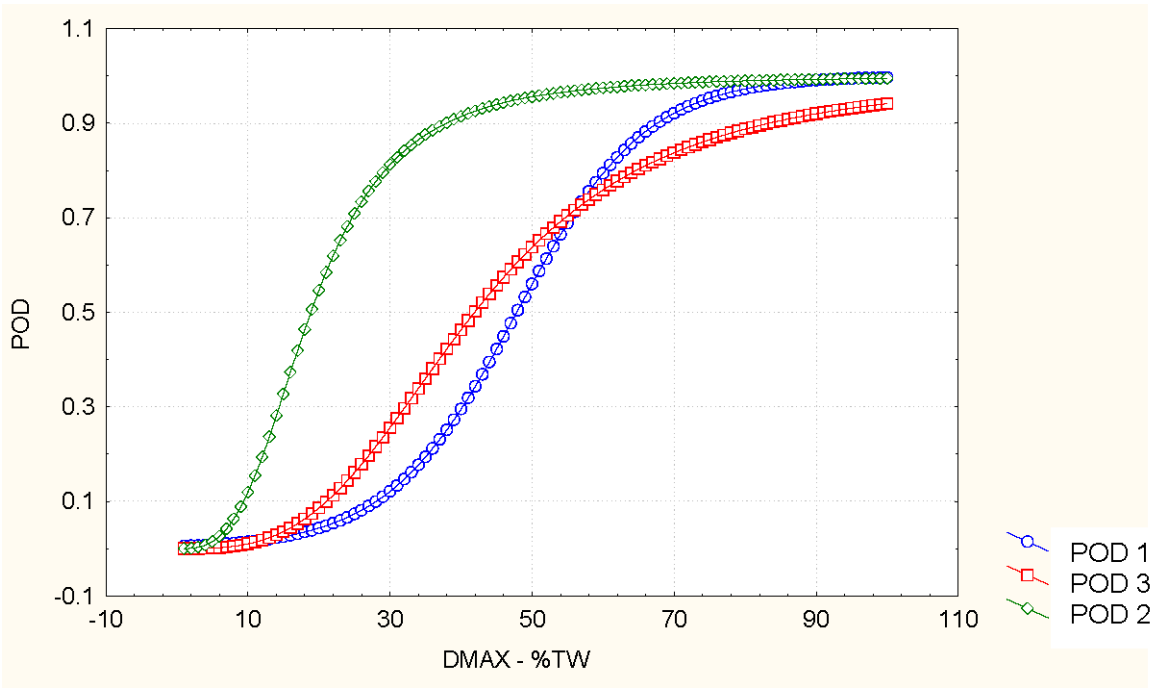

Figure 2-1
POD Function Used in Study

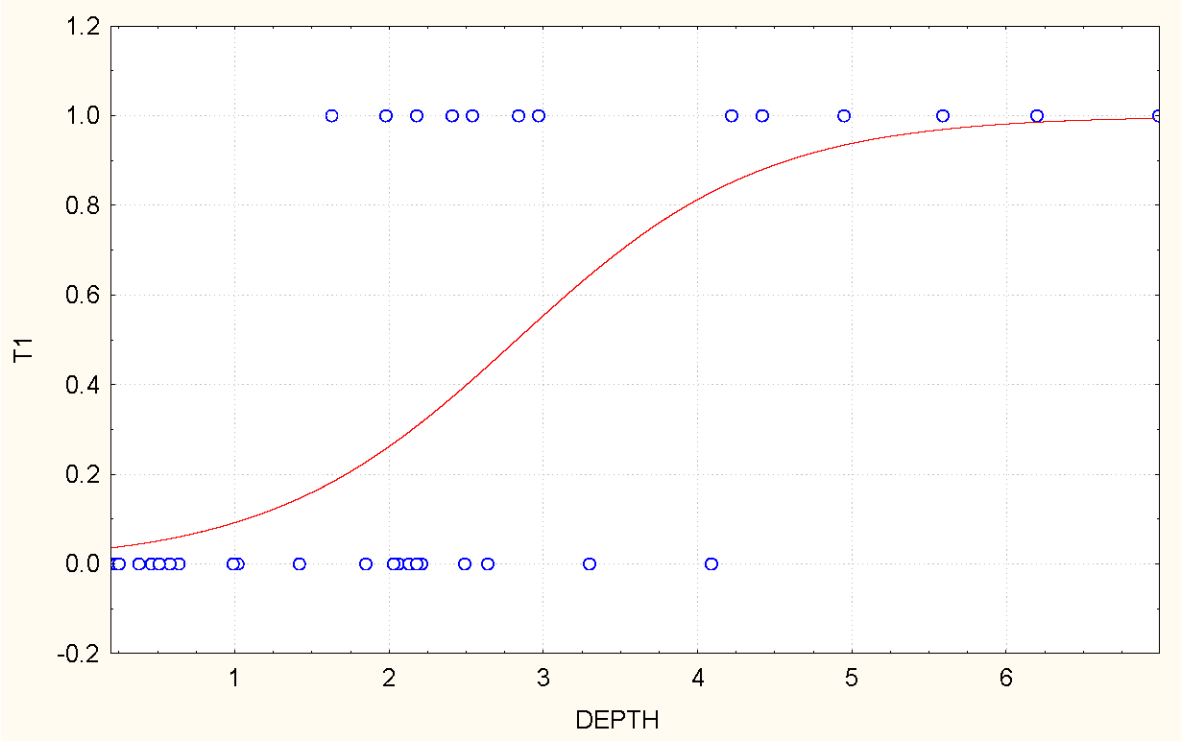

Figure 2-2
Single Team Logistic Regression

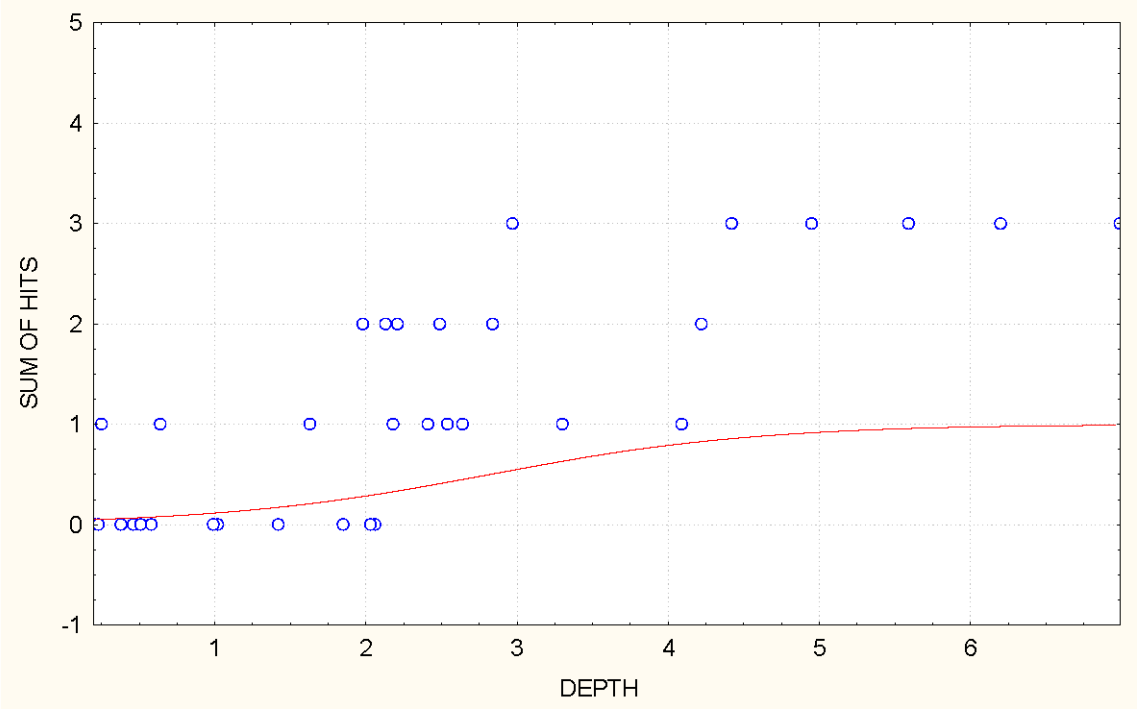

Figure 2-3
Multiple Team Logistic Regression

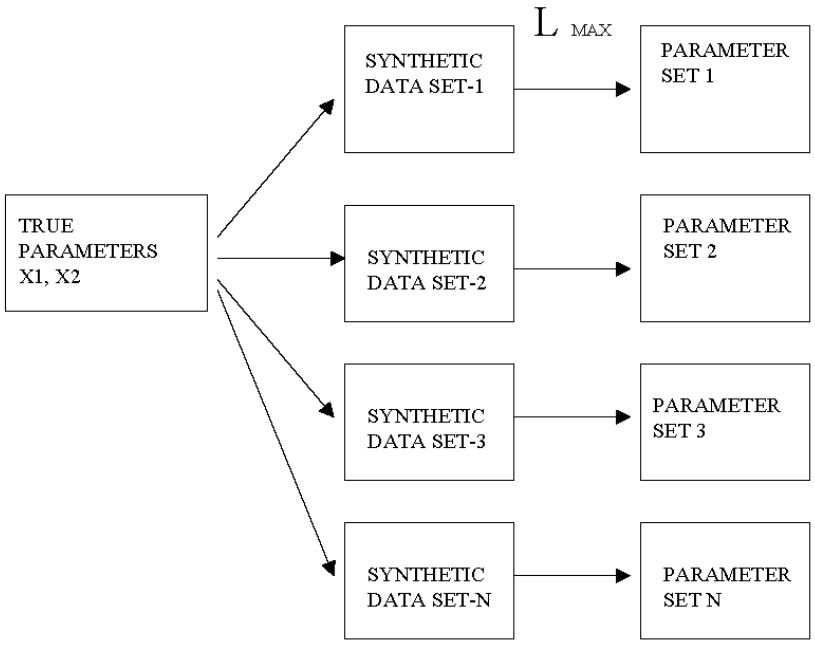

Figure 2-4
Bootstrap Monte-Carlo Process

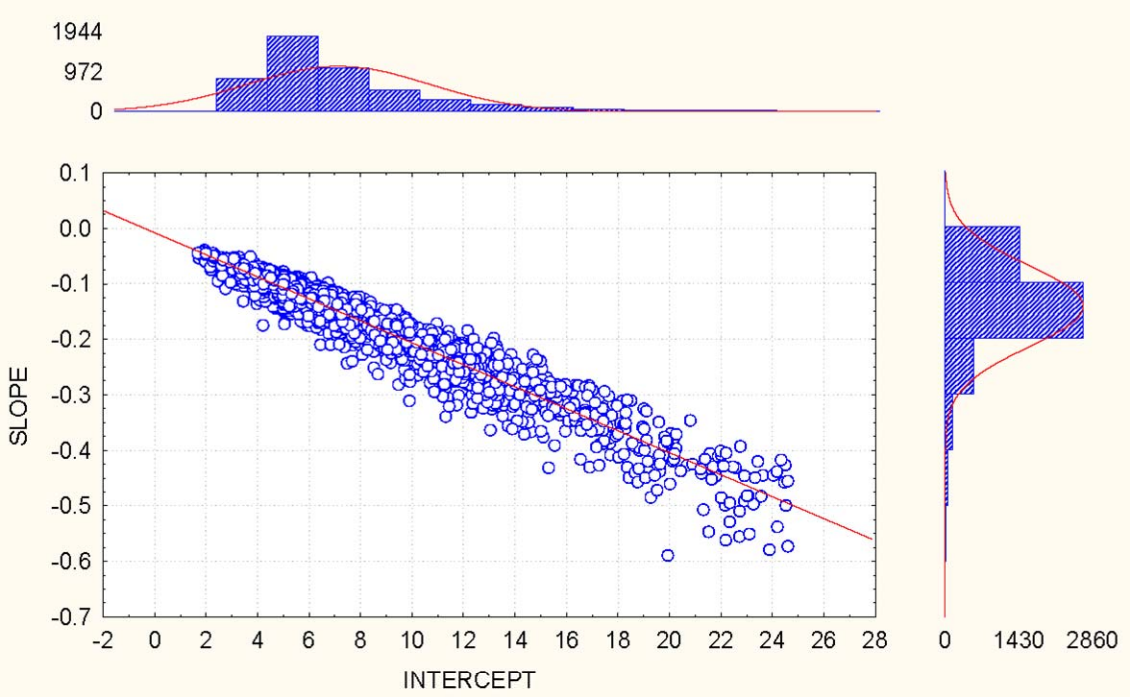

Figure 2-5
Marginal Distributions for Logistic POD Parameters (NDAT-30, NTEAM = 1)

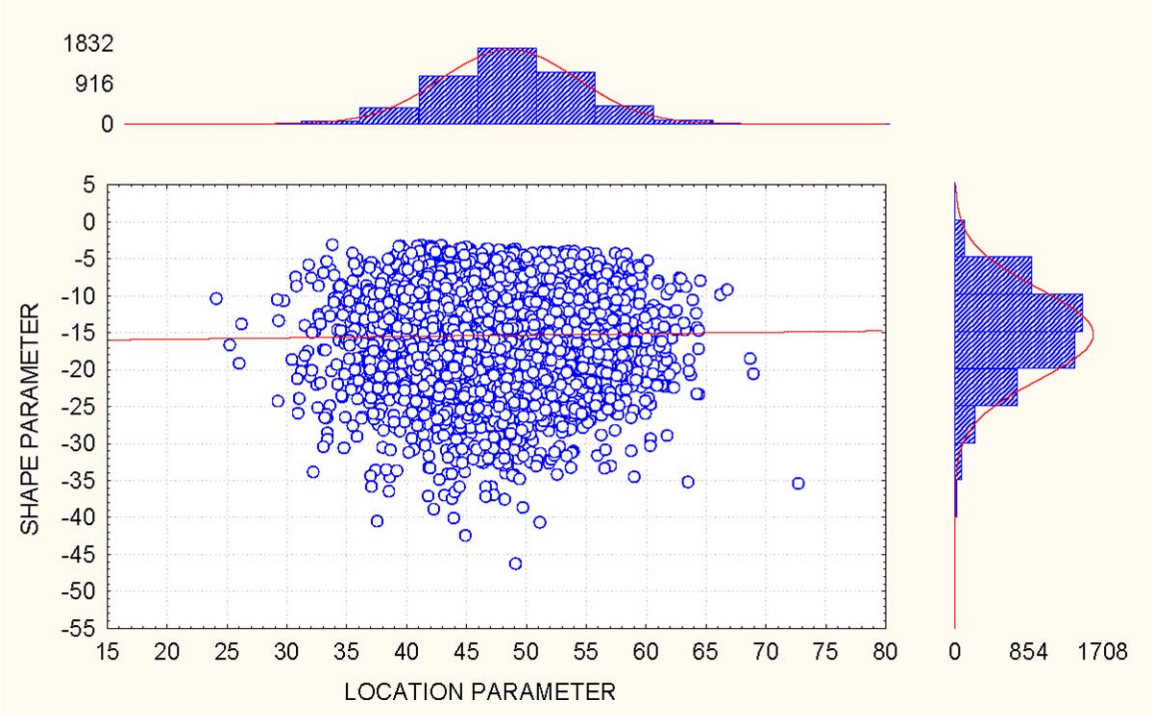

Figure 2-6
Marginal Distributions for Alternative POD Parameters (NDAT-30, NTEAM = 1)

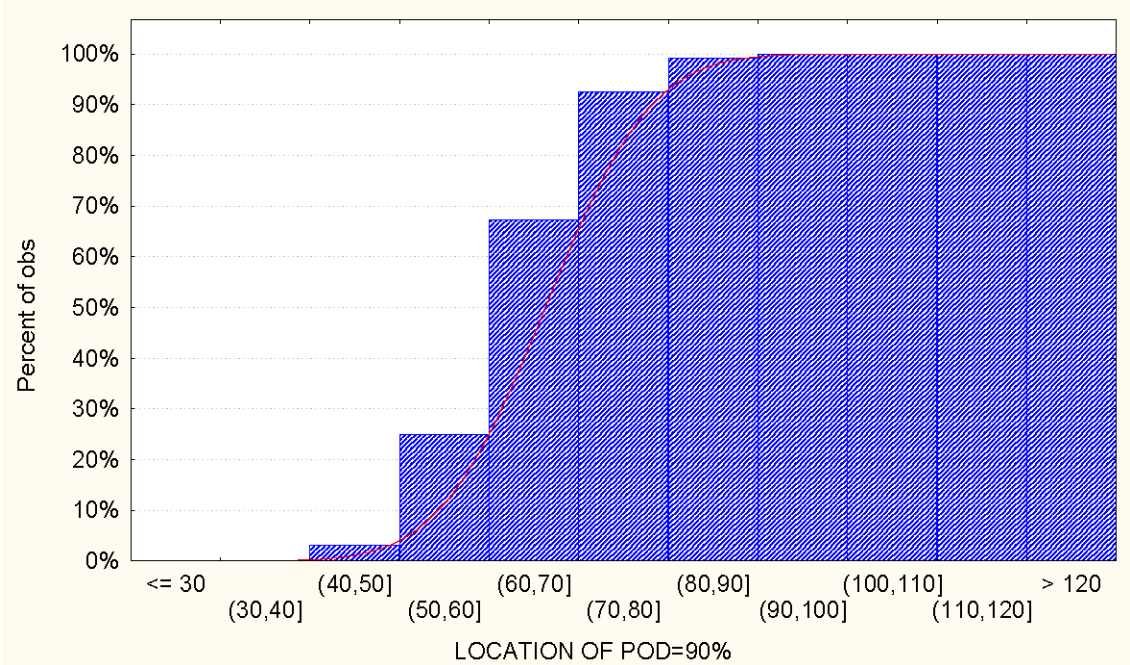

Figure 2-7
Distribution of 90% POD in % Throughwall (NDAT-30, NTEAM = 1)

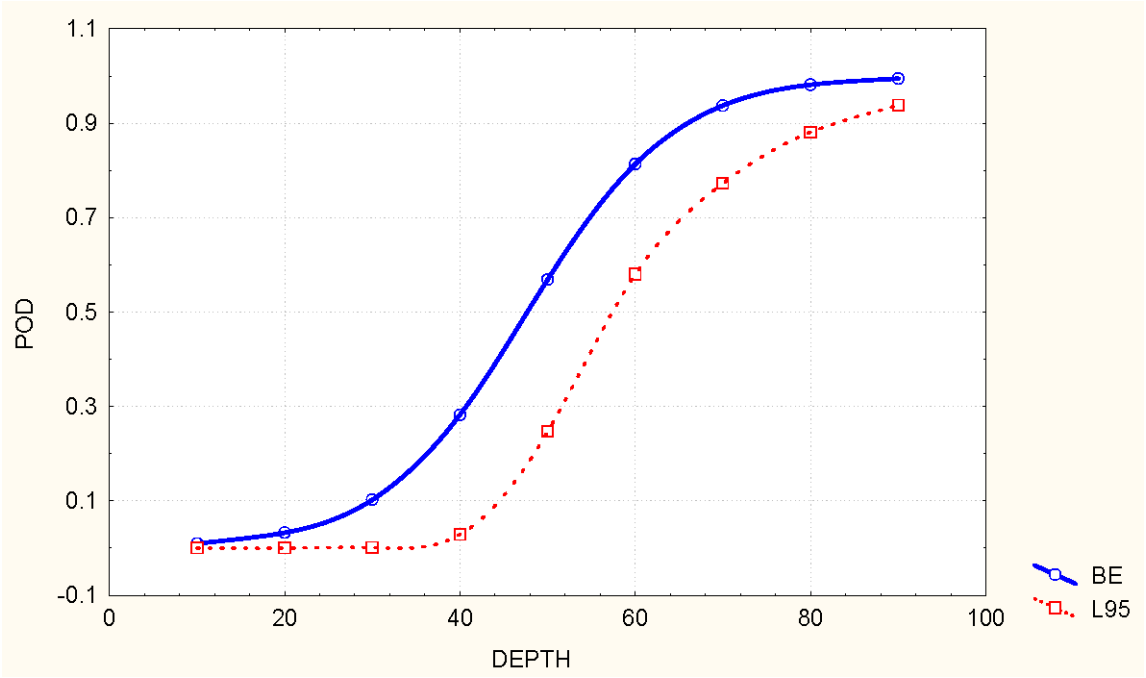

Figure 2-8
Comparison of Best Estimate POD and Lower 95% POD (NDAT-30, NTEAM = 1)

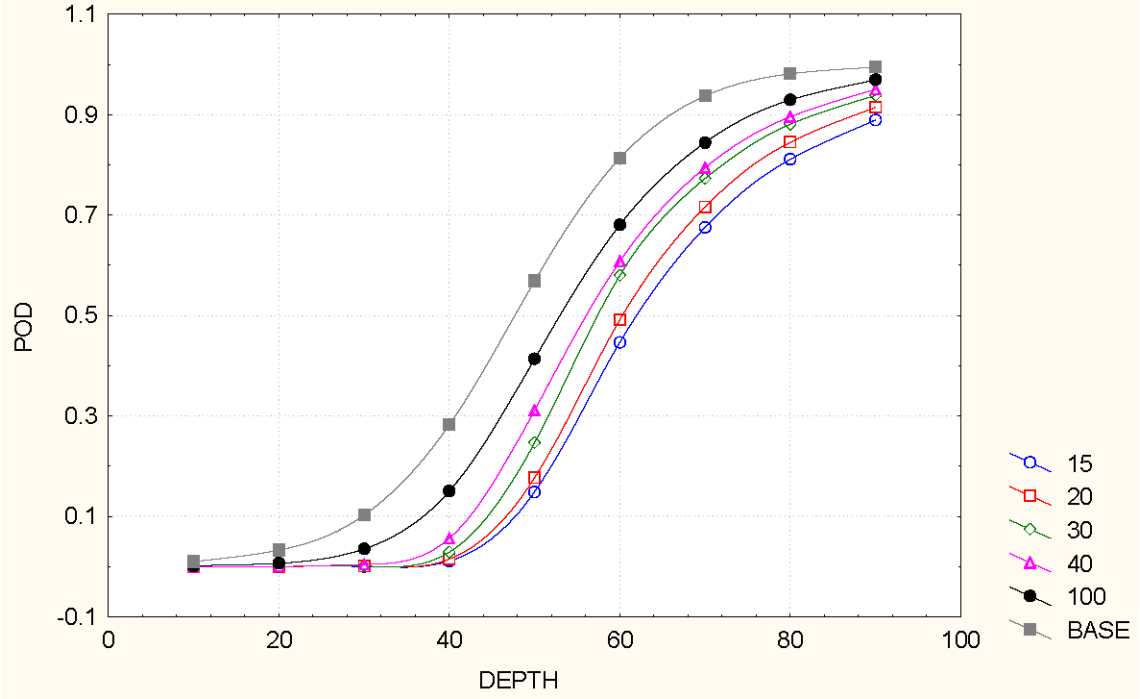

Figure 2-9
Confidence Bounds as Function of Sample Size

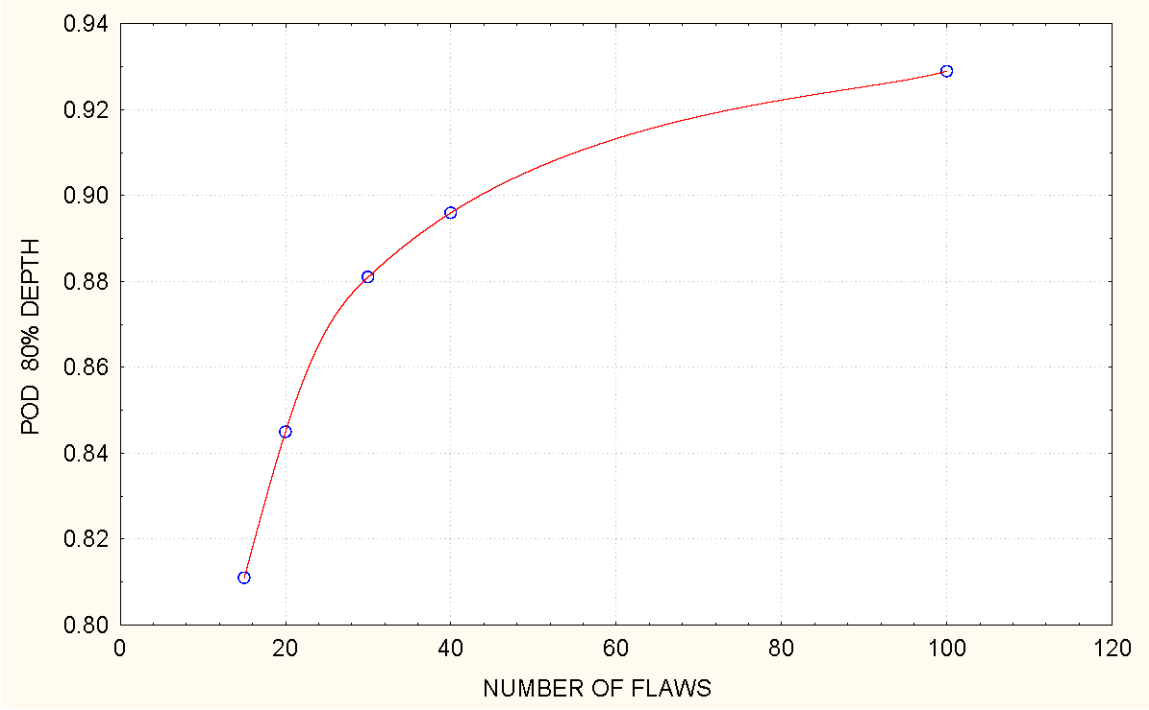

Figure 2-10
Lower Confidence Limit POD versus Sample Size (POD-1)

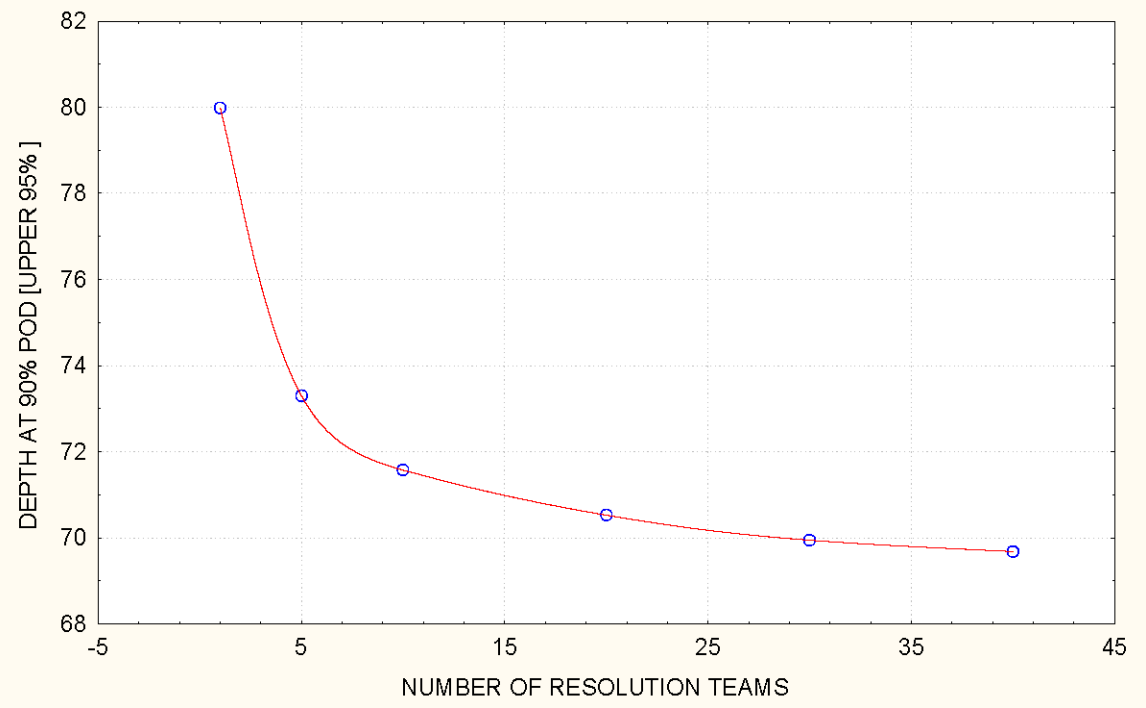

Figure 2-11
Effect of Number of Resolution Teams (POD-1)

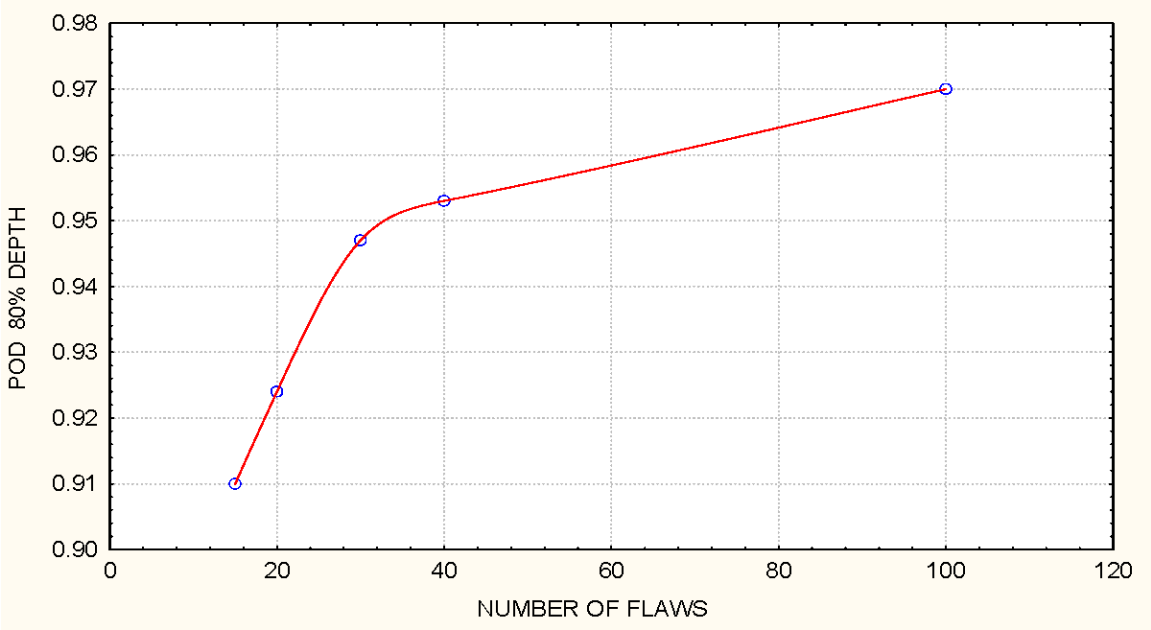

Figure 2-12
Lower Confidence Limit POD versus Sample Size (POD-2)

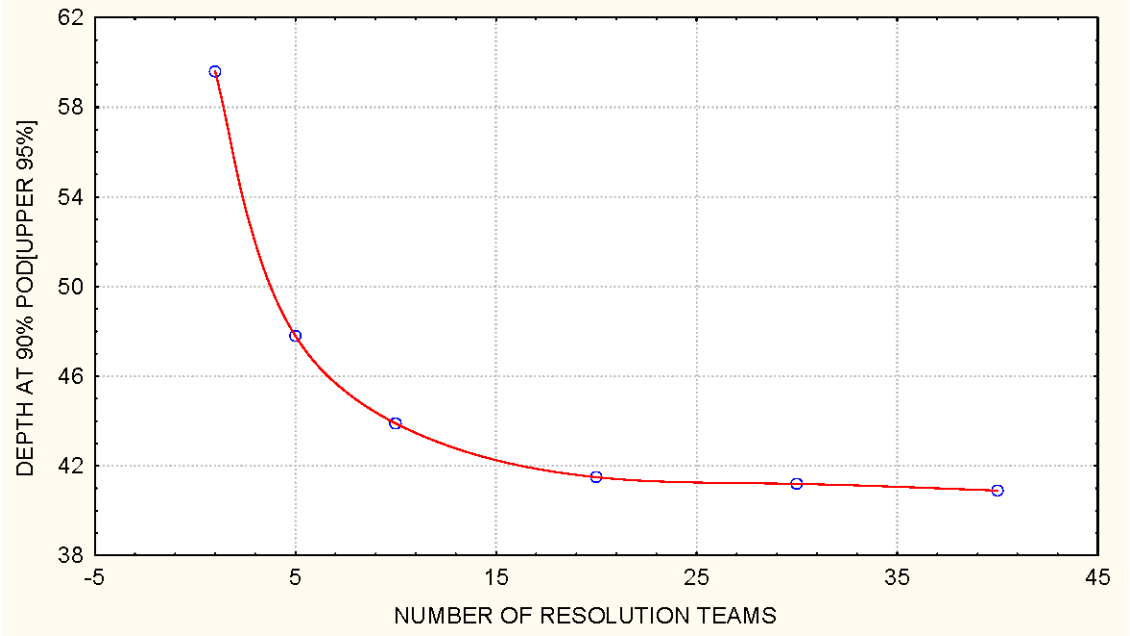

Figure 2-13
Effect of Number of Resolution Teams (POD-2)

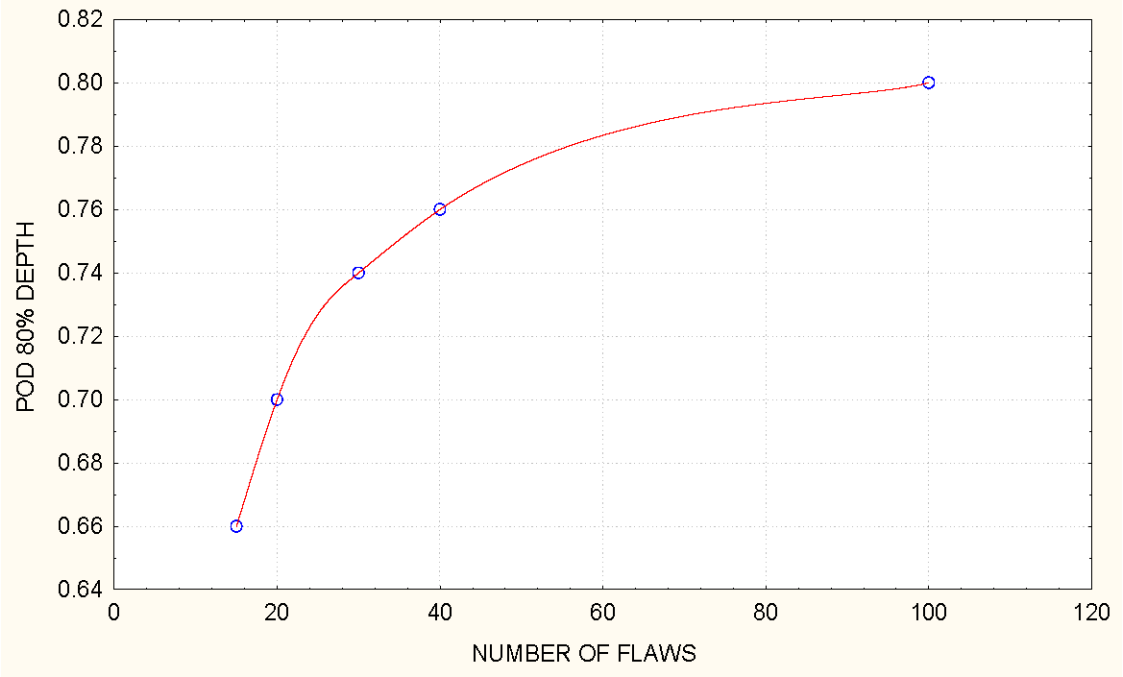

Figure 2-14
Lower Confidence Limit POD versus Sample Size (POD-3)

3
NUMBER OF UNDETECTED FLAWED SPECIMENS FOR DETECTION TESTING

Undetected but flawed specimens are considered to be too small to be expected to be detected, e.g., flaws in the 0 to 20% range, but the inclusion of such flaws is considered to be necessary to anchor the low end of the POD curve. In some cases, these small flaws show no signal response such that detection would not occur and in some cases the signal is so small that the likelihood of detection is expected to be negligible. However, care must be taken regarding the number of such indications that are added to the database due to the mathematical nature of the log-logistic distribution function that is commonly used for the POD function. The low end of the log-logistic function always has an intercept at a POD of zero for a zero value of the physical dimension. However, adding an excessive number of flawed and undetected specimens is not desirable because of the symmetry of the log-logistic function (in log space). The same is true regarding adding an excessive number of significantly flawed and detected indications. For example, the log-logistic function for POD, P, as a function of relative depth, h, would be given by,

$$\ln\left(\frac{P}{1-P}\right) = a_0 + a_1 \log h, \tag{1}$$

where the a coefficients are obtained from a regression analysis. The function is referred to as the log-odds function because the ratio of P divided by 1-P is the odds of the event happening, i.e., the probability of the occurrence divided by the probability of non-occurrence. The value of h given by,

$$h = \log^{-1}\left(-\frac{a_0}{a_1}\right) \tag{2}$$

has a POD of 0.5 and is the inflection point in log-space. The effect of adding undetected defect results to the database is illustrated on Figure 3-1. The solid squares are the undetected but real degradation that has been added to the model to anchor the low end of the curve. The solid line represents the curve fit results before adding the undetected indications and the dashed line the curve after adding the data. The net result is that the curve reflects a lower probability of detection at shallow depths as desired. However, the curve also reflects a higher probability of detection for degradation depths greater than 30% throughwall even though no additional detected indications have been added to the database. A similar effect occurs when detected significant degradation is added to the database. As shown on Figure 3-2, the POD for small indications decreases when detected large indications are added to the database.

Number of Undetected Flawed Specimens for Detection Testing

The conclusion of the above discussion is that undetected flaw POD information is not to be indiscriminately added to the POD database. The number of indications below the accepted threshold of detection should not exceed 10% of the above threshold database for that mode of degradation. Similarly, to control the high POD curve, the number of detected data points at the high POD end above the highest depth non-detected indication in the database should not exceed about 15% of the database.

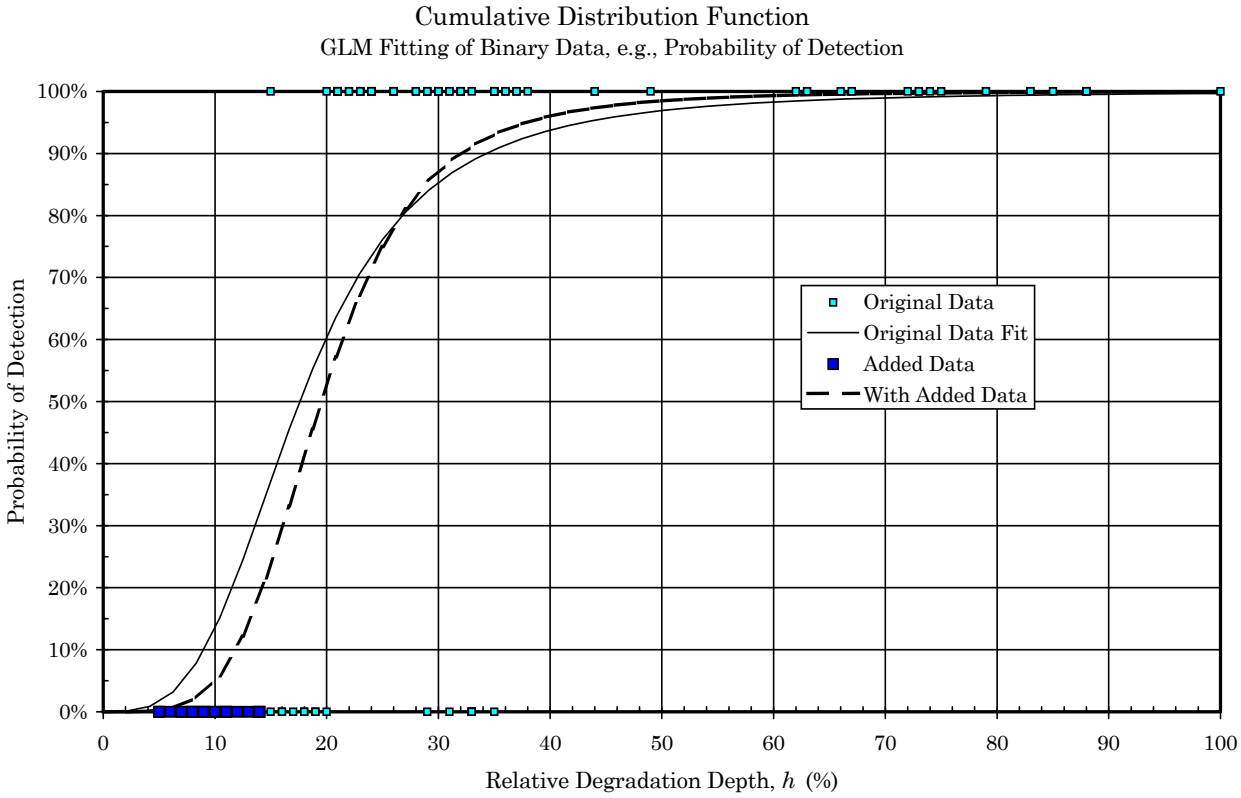

Figure 3-1
Example of Adding Non-Detected Degradation

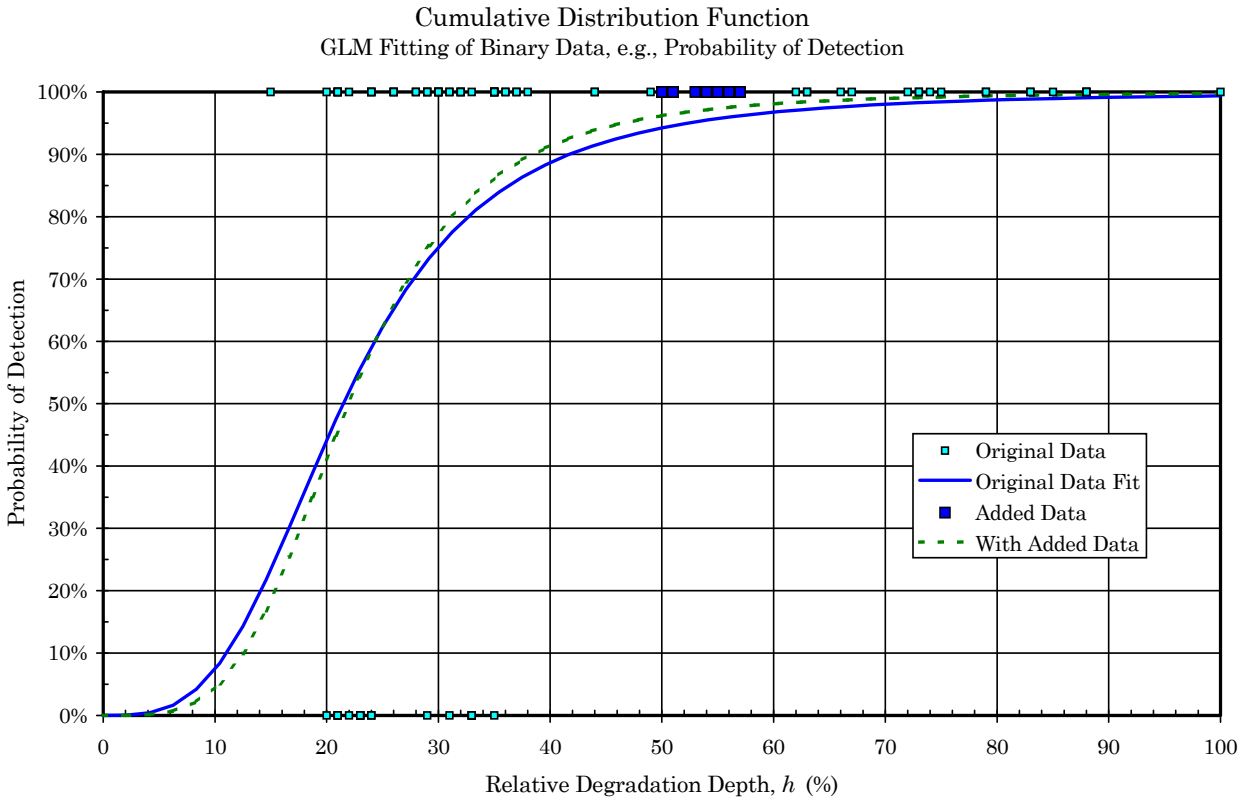

Figure 3-2
Example of Adding Biased Detected Degradation

4
NUMBER OF NON-FLAWED NDD SPECIMENS FOR FALSE CALL CONSIDERATIONS IN DETECTION TESTING

The objective of this section is to describe the basis and define the minimum number of NDD grading units to be included in the POD performance demonstration data set for assessments of false call rates. If depth is used as the single explanatory variable for POD model development, then an NDD grading unit is defined as one in which the metallographic depth of the specimen from which the grading unit "signal" is derived is zero percent.

Appendix H of Reference 1 currently does not require the inclusion of any NDD grading units for development of technique POD. Appendix G of Reference 1 requires that the number of NDDs be twice that of the non-NDD grading units. No basis for the requirement is given. The acceptable false call rate for both Appendix G and the EPRI QDA program is 10%. Other researchers (References 2, 3) state that the number of NDD grading units range should range from twice to three times the number of non-NDD grading units. Again, no explicit basis for this recommendation is given and no limiting false call rate is provided. Furthermore, Reference 3 states that for Hit/Miss POD model development, the total number of flawed grading units should be at least 60 in number. This estimate is based on experience with no analytical foundation provided. The times three factor for NDD, would imply that the total number of NDDs be $\geq$ 180.

Historically, both NDD and flawed grading units along with pass/fail criteria have been included in NDE analyst testing as a test design strategy to preclude testmanship. Inclusion of NDDs prevents an individual from passing a POD test or improving POD results by simply guessing. If the number of NDDs is larger than the number of flawed grading units, then the single trial probability of correctly reporting a hit is given by the ratio of Σ(Flawed grading units)/Σ(NDD + Flawed grading units). For example, choosing the number of NDDs to twice that or greater than the number of non-NDDs, reduces the single trial probability of correctly guessing a hit to $\leq$ 1/3. Thus the odds are in favor of the test examiner rather than the student i.e., analyst, if guessing is attempted.

The approach adopted herein is to treat the true or population overcall rate as a desired test outcome. A *sample* of NDD grading units (%TW equal to 0%) is included in the POD test that an analyst must address. The outcomes of these trials are then used to provide an upper bound estimate of the *population* false call rate. Since POD and false call rate are generally coupled, knowledge of both can be used to assess overall analyst performance.

The basic design strategy is illustrated with the data shown in Figure 4-1, which shows the *minimum* number of NDD grading units required as a function of the population false call rate

assuming zero misses. For zero misses, an individual analyst must report *all* the minimum number of NDDs as NDD i.e., no failures or false calls. The population false call rate is estimated using a one-sided upper bound confidence limit for a binomial distribution. The figure shows plots for 90% and 95% confidence limits. For the current Appendix G of Reference 1 confidence limit of 90%, the figure shows that a minimum number of 11 NDD grading units are required for a population 20% false call rate. If a single miss or failure is permitted, then for the same population false call rate, the minimum number of required NDDs increases to 18 for a 90% CL. The number of NDDs included in the performance testing can be increased to allow an increased number of false calls.

For application to the EPRI Tools program performance testing the following approach is recommended:

- Apply a false call criterion to prevent excessive conservatism during the performance testing. This criterion will be reevaluated after the initial test results are reviewed.
- The false call criterion should be applied to the analysis team as a whole i.e., resolution false call rate. Accordingly, no false call limit for production analysts but rather a limit for resolution analysts. This would also permit production analysts to be conservative in their initial reporting which they should be and usually are.
- Choose a higher false call rate for bobbin coil detection since it is often used as a screening tool for further confirmation and diagnosis by rotating probe.
- Accordingly, increase the current Appendix G false call rate limit for bobbin coil detection based on resolution analysis to 20%. Maintain the current Appendix G false call rate of 10% for rotating probe analysis; again this is a ream or resolution analysis value with no limit n production analysis.
- If the limits described above are exceeded during performance testing, the team results will be rejected.
- The least impact on sample number requirements is achieved using a zero failure or miss criterion requiring a minimum of 11 NDD grading units at 90%CL.

Non-flawed or NDD specimens are to be included in performance testing for POD development to provide a constraint on the NDE analysts against overly conservative false call rates. The NDD requirements are established to obtain a 90% confidence on an acceptable false call rate, which can be used to define an acceptable number of false calls for a given NDD population size. For bobbin coil analyses, conservative calls are to be encouraged to enhance the POD when rotating coils are to be used for flaw confirmation. RPC analyses should be less conservative than bobbin analyses since overcalls can lead to unnecessary tube repair. Since the primary and secondary analysts are encouraged to provide conservative calls, the false call requirements are applied to the results of the resolution analyst.

The false call requirements are ≤20% at 90% confidence for bobbin detection and ≤10% at 90% confidence for rotating probe detection. Dependent on the difficulties in obtaining NDD specimens, the number of specimens can be increased to provide increased allowances for false calls. The false call requirements are applied to the results of the resolution analyst. If the false call rate for a resolution analyst's team exceeds the acceptance limits, the detection results for

that team are not included in the POD evaluation for the associated dataset or ETSS. The minimum number of NDD specimens with no false calls permitted would be 11 specimens for bobbin coil detection and 22 specimens for rotating probe detection. An allowance for at least one false call is suggested but not required.

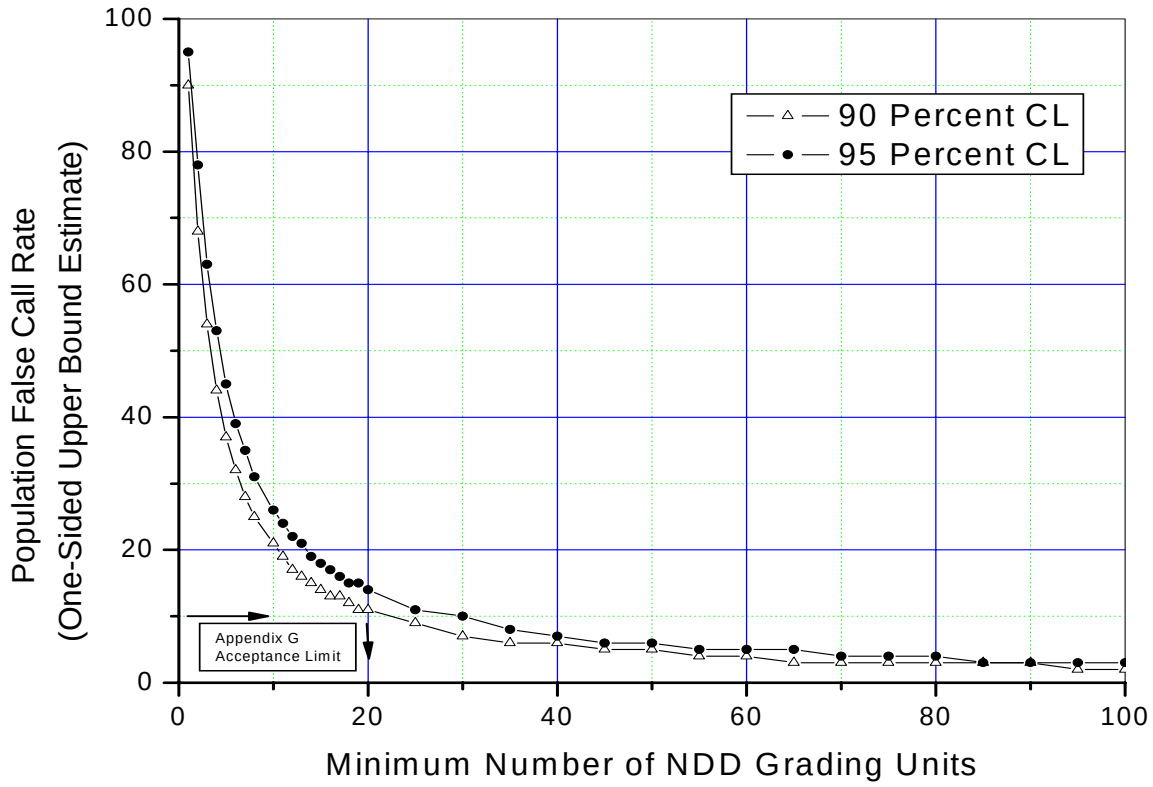

Figure 4-1
Minimum Number of Required NDD Grading Units as a Function of Population False Call Rate (Plot assumes zero misses i.e., none of the grading units is reported incorrectly)

5
NUMBER AND DEPTH DISTRIBUTION OF SPECIMENS FOR NDE SIZING TESTING

The end product for developing NDE sizing uncertainties is a regression equation relating the actual depth of the indication to the predicted depth from NDE sizing analyses. There are a number of assumptions that are made regarding the analyses to be performed. These are provided in the following list.

1. The relation will be expressed as a linear, 1st order equation.
2. The intent of the testing is to minimize the error of prediction for a number of physical values simultaneously.

There are a number of factors that influence the width of the zone into which future values of the destructive examination (DE) measurements are expected to fall. These are the parameters of the regression equation and the standard error of the regression predictions along with statistical distribution values, F and χ^2, that are a function of the number of data pairs used in the analysis. As the number of data increase, the influence of the F and χ^2 distribution terms decreases.

A one-sided tolerance bound for DE depths as a function of NDE depths would be expressed as,

$$h_{DE} \le b_0 + b_1 h_{NDE} + k\, s\,, \tag{3}$$

where the b coefficients are those that result from the regression analysis from n pairs of data, s is the standard error of the regression and k is referred to as the tolerance factor. The value of k is a function of the F and χ^2 distributions and decreases with increasing sample size as noted in the preceding paragraph. The effect of sample size is illustrated on Figure 5-1 where the F and χ^2 distribution terms are shown along with the value of the tolerance factor. It is apparent from the figure that the tolerance factor diminishes rapidly with sample size up to about an n of 20 to 30. Thereafter, the asymptotic value is approached slowly. Both the F and χ^2 terms diminish rapidly in the same range. Examination of the figure shows that the rate is governed more strongly by the χ^2 term for the standard deviation. This is not unexpected. The visual conclusion to be drawn from the figure is that the number of data pairs should be greater than 30 and does not need to be more than about 60 to 70. Engineering judgment indicates that a range of 30 to 40 samples is appropriate. From Figure 5-1, the value drops by about 8% between 30 and 100 data pairs, which is not seen as significant relative to the effort to develop and examine an additional 60 specimens. The reduction from increasing the number of specimens from 60 to 100 would be about 2%.

Number and Depth Distribution of Specimens for NDE Sizing Testing

The required minimum number of samples for a sizing correlation depends on the intended application for the correlation. For NDE sizing uncertainties that are to be used as a basis for leaving indications smaller than the repair limit in service, the correlation must be established with a high degree of confidence and the recommended number of samples for NDE performance testing is a minimum of 30 samples. If 30 specimens are not available, the minimum number of required specimens is 20, which represents the lower uncertainty end of the knee in the curves of Figure 5-1. If < 30 samples are available for performance testing, the correlation is limited to support of tube integrity analyses, and the ETSS should identify that the correlation is not adequate for sizing indications to leave indications in service.

The other issue to be addressed is the range and spread of the sizes to be categorized. For example, if a regression analysis relating depths is to be performed, it is unlikely that a smooth distribution of specimens will be found. Hence, requirements are usually stipulated that restrict the number of data within subranges or bins. For a linear first order relation for the sizing correlation, the ideal distribution for the independent variable is to have half at each extreme. This has the effect of minimizing the variance of the slope coefficient found by the regression analysis. However, it is worth having data in the middle of the range for the purpose of detecting whether or not there are anomalies in the data evaluation process. Thus, it is recommended that the data be split about evenly in three ranges such as depth ranges of 0 to 35%, 36% to 65%, and 66% to 100%. This is meant to address the range of levels of depth to be investigated, the extremes to be investigated, how the data should be spaced, and the number of observations in each range of depth.

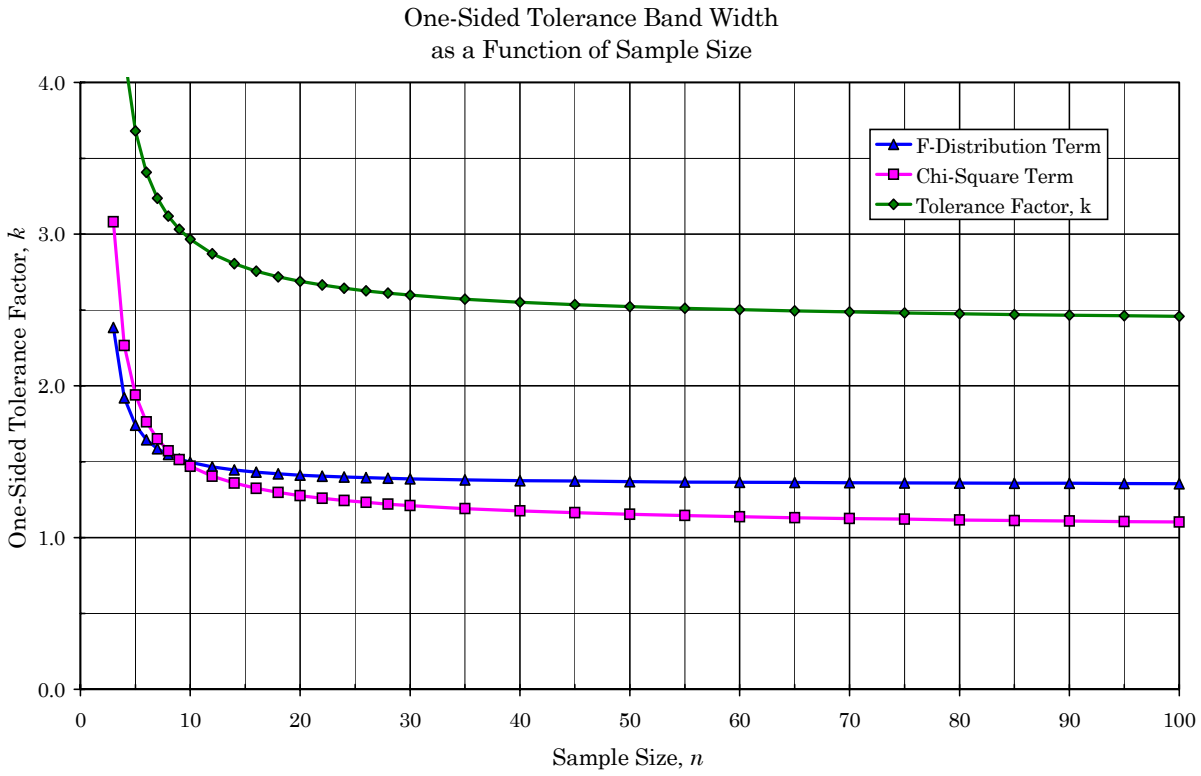

Figure 5-1
Effect of Sample Size on Tolerance Interval Width

6
NUMBER OF NDE ANALYSTS FOR SIZING TESTING

The issue of concern in terms of the number of required analysts for NDE sizing is the analyst variability component. This must be included in both Operational Assessment and Condition Monitoring work. A sufficient number of analysts are required to obtain a precise estimate of analyst variability, which can be expressed as a standard deviation. The confidence interval for analyst standard deviation can be expressed as a function of the Chi-squared distribution function and N (sample size) as described in Section 4 for the number of specimens. The confidence bounds on the standard deviation are non-symmetric and can be expressed as a sample size dependent set of coefficients by which the sample standard deviation can be multiplied. These are shown in Figure 6-1. As can be seen from the figure, the more critical upper bound decreases greatly prior to a sample size of 10 to15 analysts. Engineering judgment suggests that a prudent sample size would be ≥ 10. The results of Figure 6-1 apply to the number of NDE sizing teams, which typically include a sizing analyst and an independent technical reviewer (ITR). The number of analysis teams for detection testing is recommended to be ≥ 10 for a generic ETSS program. If the performance testing is being conducted for a site specific POD based on testing of the analysts used in the inspection, the number of analysis teams should be a minimum of 5.

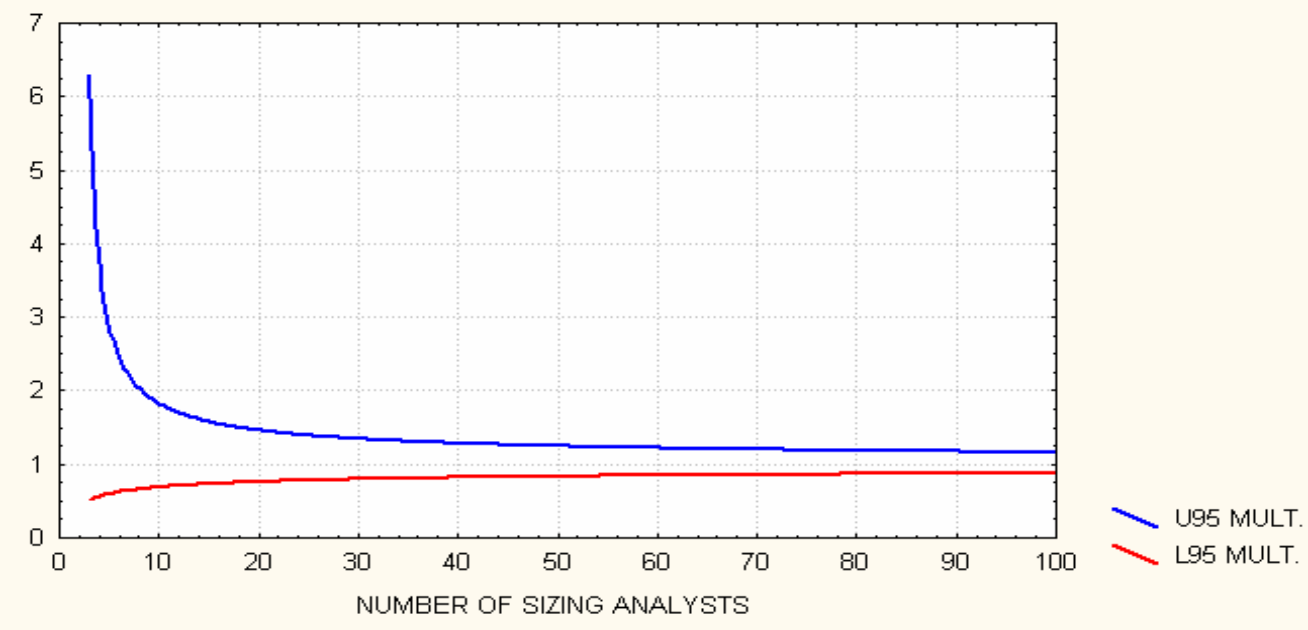

Figure 6-1
Upper and Lower Confidence Limits (shown as multiplier of sizing analyst standard deviation)

7
REASSESSMENT OF NUMBER OF TEAMS REQUIRED FOR DETECTION TESTING

This section provides a reassessment of Section 2.0 for the number of NDE analysis teams required for POD testing. The results of Section 2.5 (e.g., Figure 2-11) are based on statistical analyses for the influence of uncertainties at 95% confidence for varying number of teams on the depth for a given POD of 0.90. This section uses the results of completed 10 team performance testing to assess the sensitivity of the PODs to a reduced number of analysis teams. The influence of uncertainties on the POD at a given depth and on the depth at a given POD is addressed in this section.

7.1 General Considerations on Influence of Number of Analysis Teams on POD Distributions

Lower confidence PODs decrease the POD at a given depth and increase the depth at a given POD. Tube integrity Monte Carlo analyses are dependent on the change in POD for a given depth since the POD is applied to flaws of a specified depth. Deterministic analyses for tube integrity can be based on applying a specified POD value at the lower 95% confidence value (e.g., POD of 0.95 at lower 95% confidence depth). The uncertainty on depth for a high POD value can be very large since the POD curve is tending toward a very small slope. Figure 7-1 shows an example of this effect for which a POD of 0.80 shows an increase in depth from about 60% for the nominal POD to about 80% for the lower 95% confidence POD. This effect leads to substantial conservatism in deterministic analyses applying a high POD value at a high confidence level. Consequently, for this deterministic analysis method, the principal concern relative to selecting the number of analysis teams is the influence of the number of teams on the magnitude of the POD uncertainty.

The influence of POD uncertainties in Monte Carlo analyses is very small. PODs are applied to specific depth flaws so that the smaller uncertainty on POD at a given depth is applied. More significantly, Monte Carlo sampling with POD uncertainties leads to some SG samples with more indications (lower POD sample) and about the same number of SG samples with fewer indications (higher POD sample). The effects of higher or lower POD samples tends to average to negligible influence on probability of burst or leak rate analyses since they average to the number of indications obtained from the nominal POD distribution. The number of indications only multiplies the single indication burst pressures and leak rates so that the SG burst probabilities and total SG leak rates converge toward that obtained from the application of the nominal POD. For POD distributions approaching 0 and 1 with significant uncertainties, the Monte Carlo process tends to modify the effective average POD. This results since symmetric uncertainties cannot be applied on the low end near 0 or the high end near 1 due to the cut off limits at 0 and 1. The net effect is to slightly lower the average POD near 1 and increase the

average POD near 0. For Monte Carlo analyses, the principal concern relative to selecting the number of analysis teams is the influence of the number of teams on the nominal POD values.

Figures 7-1 and 7-2 show the effect of reducing POD uncertainties by increasing the number of analysis teams from 1 to 10. For 10 teams, the lower 95% confidence POD is a small reduction relative to the nominal POD. Figure 7-3 shows the bobbin depth increase for selected POD values at the lower 95% confidence relative to the nominal POD as a function of the number of analysis teams. The largest depth increases for a fixed POD at the 95% confidence level occur at the highest POD values where the POD tends to flatten out with a small slope. The results of Figure 7-3 are comparable to the prior assessments in Figures 2-11 and 2-13 except that the prior results are plotted as POD values rather than the relative error of Figure 7-3. As noted above, this uncertainty significantly adds to conservatism for deterministic analyses applying a high POD at a high confidence level to define the largest potential indication left in service.

Figure 7-4, shows the reduction in POD at selected depth values as a function of the number of analysis teams. The POD reductions are small and approximately independent of depth. This figure represents the POD uncertainty effect when applying Monte Carlo analyses.

7.2 POD Sensitivity Based on Evaluation of 10 Team POD Performance Test Results

Performance testing for 10 NDE analysis teams has been completed under the EPRI Tools for Tube Integrity Program for axial ODSCC. The difficulties in organizing 10 teams for testing and the high cost of this extensive testing are the primary reasons for the reassessment of the number of required analysis teams in this report. The 10 team results permit an assessment of the POD sensitivity to the number of teams. The individual team detection results for the 10 teams were applied to form 3 Groups of 3, 5 and 7 teams. The groupings formed from the 10 team results are given in Table 7-1, where each letter represents one of the 10 teams. The detection results were then applied to obtain a POD distribution for each of the groups. Since the objective of this assessment is to evaluate the impact of reducing the required number of teams to less than 10, the POD errors for each sub-group were defined as differences from the 10 team results. This definition of error is equivalent to defining the 10 team results as “truth” and directly shows the effect of reducing the number of analysis teams to less than 10. POD errors were assessed for the nominal POD and the lower 95% POD for the POD at a given depth. POD errors were also assessed for the increase in depth at the lower 95% confidence at a given POD value.

Figures 7-5 to 7-7 compare the 10 team nominal POD with the 3, 5 and 7 team sample PODs. The largest difference between the 10 team POD and the group samples decreases as the number of teams increases as would be expected. With the exception of only one of the 3 team results, the differences in sample nominal PODs from the 10 team result are small. Since only 3 group samples are used for each number of teams comparison, the largest variations between group samples may not have been identified. One team primarily contributes to the high POD for the 3 team group 1 result in Figure 7-5. When the number of teams is increased to 5, the influence of the one team with a high POD tends to be averaged to a modest influence on the combined POD. Three teams for POD testing do not appear to be sufficient to average out the potential for a team tending toward outlier results.

Figures 7-8 and 7-9 show the POD error relative to the 10 team results as a function of the number of analysis teams at depths of 40%, 70% and 95%. The Figure 7-8 data reflect the differences between the nominal POD curves of Figures 7-5 to 7-7. The trend lines shown in these and subsequent figures are provided only to show the overall trend of fitting the data points. The extremes for the errors are more important than the trend line for assessing the number of analysis teams. The data in Figure 7-6 are differences between the lower confidence values for N teams and the 10 team results and are influenced by differences in the nominal PODs. For example, a 5 team result with a higher nominal POD may have very good agreement at the lower 95% values with the 10 team result since the increased uncertainty for 5 teams can be offset by the higher nominal POD. The results of Figures 7-8 and 7-9 show negligible differences in POD at a given depth when at least 5 teams are included in the POD analysis.

The nominal and lower 95% depth errors relative to the 10 team results at PODs of 0.40, 0.70 and 0.80 are shown in Figures 7-10 and 7-11. The errors increase as the POD evaluated is increased due to the lower slope of the POD curves at higher PODs. The largest nominal errors in Figure 7-10 of 8% and 4.5% at POD = 0.80 for 3 and 5 teams are significant for tube integrity applications that might apply the depth at a high POD at a high confidence level. For these samples, the lower 95% confidence errors in Figure 7-7 are smaller than can be expected since the largest nominal errors occurred for PODs higher than the 10 team result.

To assess the sensitivity of the POD errors for less than 10 teams to the POD being evaluated, the process of sampling 3, 5 and 7 teams was repeated for the +Point TSP results of the performance testing. The +Point POD results for each of the 10 teams shows less variation than obtained for the bobbin data so it can be expected that the data would show less sensitivity to the number of analysis teams used to develop the POD. Figures 7-12 to 7-14 show the comparisons of the 10 team POD to the POD results from sampling 3, 5 and 7 teams. From these figures, it is seen that only the 3 team results show any significant variation from the 10 team results. The nominal and lower 95% confidence POD errors at given depths relative to the 10 team results are shown in Figures 7-15 and 7-16. The larger nominal POD error of Figure 7-15 for 3 teams at 40% depth reflects the one team in Figure 7-12 with the highest POD. The POD errors at the lower 95% confidence in Figure 7-16 are smaller than would be obtained if one of the samples had a POD significantly lower than the 10 team POD, which can occur with equal likelihood to the results of this assessment. Overall, the +Point TSP results support the adequacy of using 5 teams to develop the performance test POD.

7.3 POD Sensitivity to Number of Analysis Teams Based on Monte Carlo Sampling

As a general check on the 3 group POD analyses of the performance test results in Section 7.2, the process was simulated by Monte Carlo analysis using 3 trials for 3, 5, 7 and 10 teams. Given the small number of random Monte Carlo trials, the variability between trials is only an example of the trends and does not reflect the full range of uncertainties obtainable with larger samples. The definition of “truth” for the Monte Carlo process is the 10 team performance test POD evaluated from fractional detections. This yields the same nominal POD as the weighted GLM analyses but has uncertainties typical of one team rather than 10 teams. One Monte Carlo POD data trial was performed by performing 10 samples of the “truth” POD with uncertainties. Each POD data trial includes 135 depth samples using discrete sampling of the performance test

depths. A sample POD is compared to a random number between 0 and 1 to define hit or miss dependent upon whether the sample POD value exceeds the random number. POD hit/miss trial results were then obtained from the 10 samples using 3, 5, 7 and 10 of the samples to represent the corresponding number of teams. These trial results were then input to GLM POD analyses to obtain trial PODs with defined uncertainties.

Figures 7-17 to 7-20 compare the nominal 10 team performance test POD with the 10, 7, 5 and 3 team Monte Carlo sample PODs as examples of the POD variability dependence on the number of teams. As seen in Figures 7-17 and 7-18, the 10 and 7 team examples include one POD team tending toward outlier behavior while 2 of the 3 trial results yielded PODs essentially identical to the 10 team performance test POD. The 5 team results of Figure 7-19 show moderate POD variability over all depths while the 3 team results of Figure 7-20 show the largest POD variability for Trial 2. The 3 team Trial 2 variability is seen to be averaged out when two teams are added to the analysis for the 5 team Trial 2 result in Figure 7-19. The trend of POD variability with the number of analysis teams obtained for these Monte Carlo trials is very similar to that obtained by grouping the performance test results (Figures 7-5 to 7-7).

7.4 Recommendation on Number of Teams for Detection Testing

Considerations for selecting the minimum number of analysis teams for POD performance testing should be the POD uncertainty at a given depth rather than the uncertainty in depth at a given POD value. The latter is applicable only to very conservative tube integrity analyses that apply a high POD and a high confidence value to bound the potential indication left in service. Revision 2 of the Steam Generator Integrity Assessment Guidelines (Reference 4) defines non-statistical analysis methods that use a nominal POD = 0.95, but do not require the application of the lower 95% confidence that was included in Revision 1 of the guidelines. Some PODs, such as the bobbin TSP example applied in this report, may not reach 0.95 even at 100% depth so that statistical methods, such as Monte Carlo, may be required. As noted above, the Monte Carlo process for tube integrity analysis, which applies POD uncertainties at given depths, results in a small influence of POD uncertainties on burst pressure and leak rate calculations.

Overall, the variation in the nominal POD with the number of analysis teams is the primary basis for defining a requirement on the minimum number of teams. The minimum number of analysis teams should be able to accommodate some detection variations between teams, such as one team tending toward outlier behavior, with minimal influence on the nominal POD. The above results show that the use of 5 teams to develop a performance test POD is adequate to accommodate team-to-team variability with minimal affect on the combined team POD.

Based on the above, it is recommended that a minimum of 5 teams be used for detection performance testing.

Table 7-1
POD Team Groups for Assessing Sensitivity to Number of Teams

Number of Teams	Group Number	Teams Comprising Group
3	1	A, C, E
	2	B, D, H
	3	F, G, I
5	1	A, C, E, G, I
	2	B, D, F, H, J
	3	B, C, F, G, J
7	1	A, B, C, E, G, H, I
	2	A, D, E, F, G, H, J
	3	B, C, D, F, G, I, J

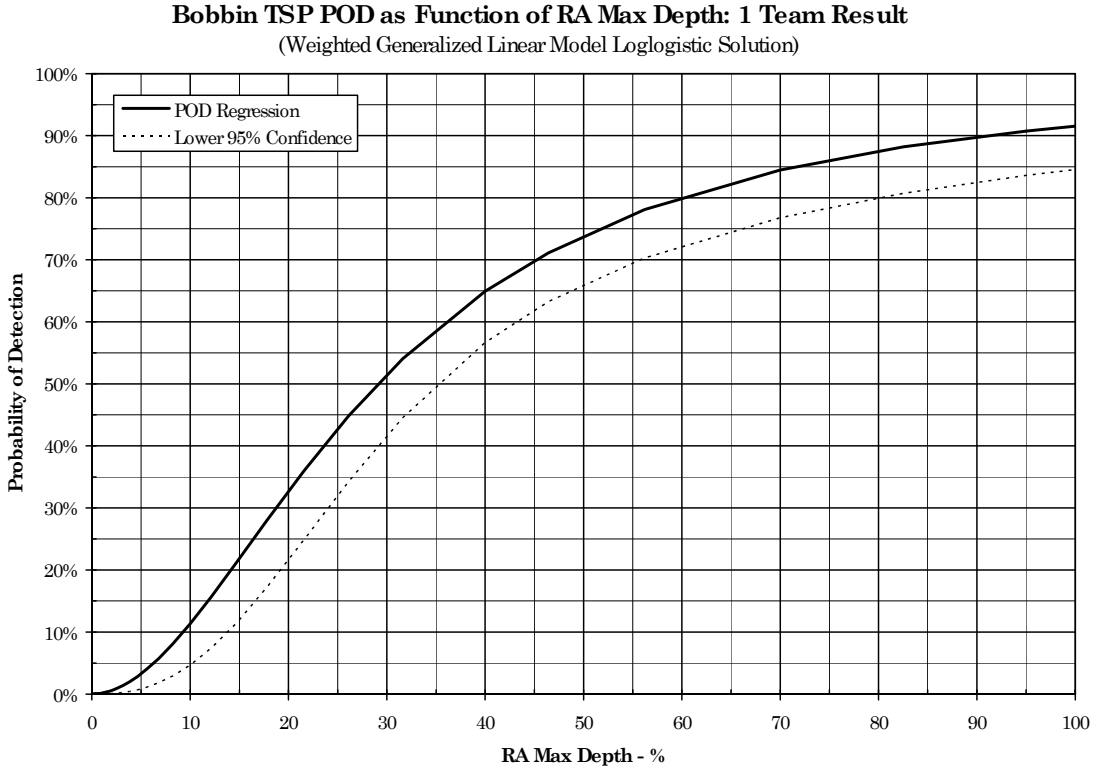

Figure 7-1
Bobbin TSP Nominal and Lower 95% Confidence POD for 1 Analysis Team

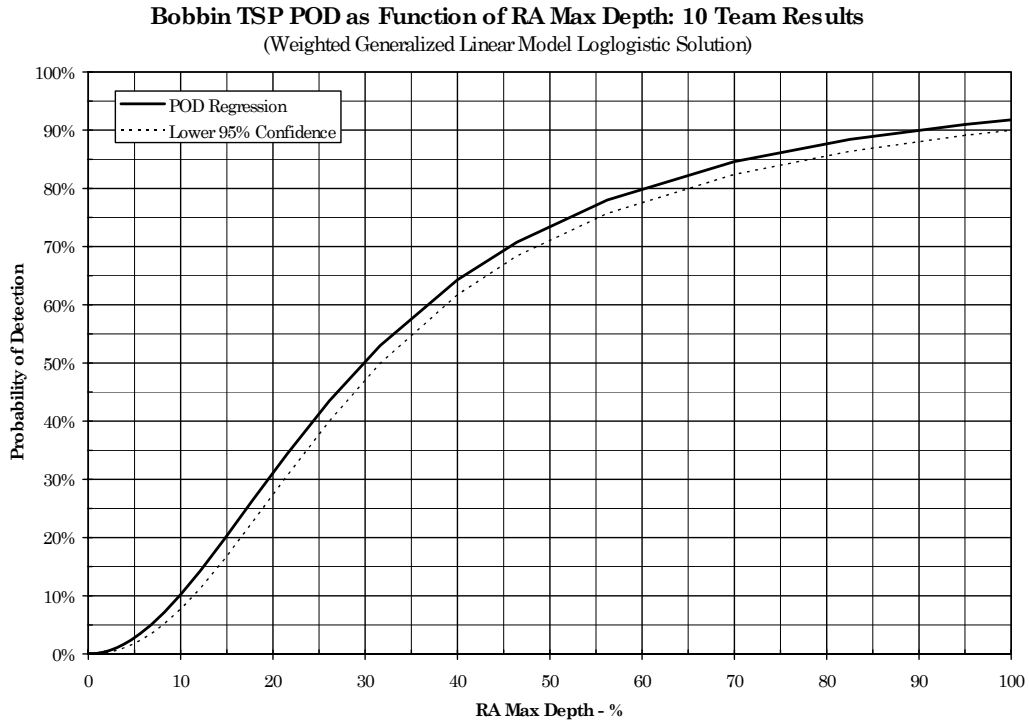

Figure 7-2
Bobbin TSP Nominal and Lower 95% Confidence POD for 10 Analysis Teams

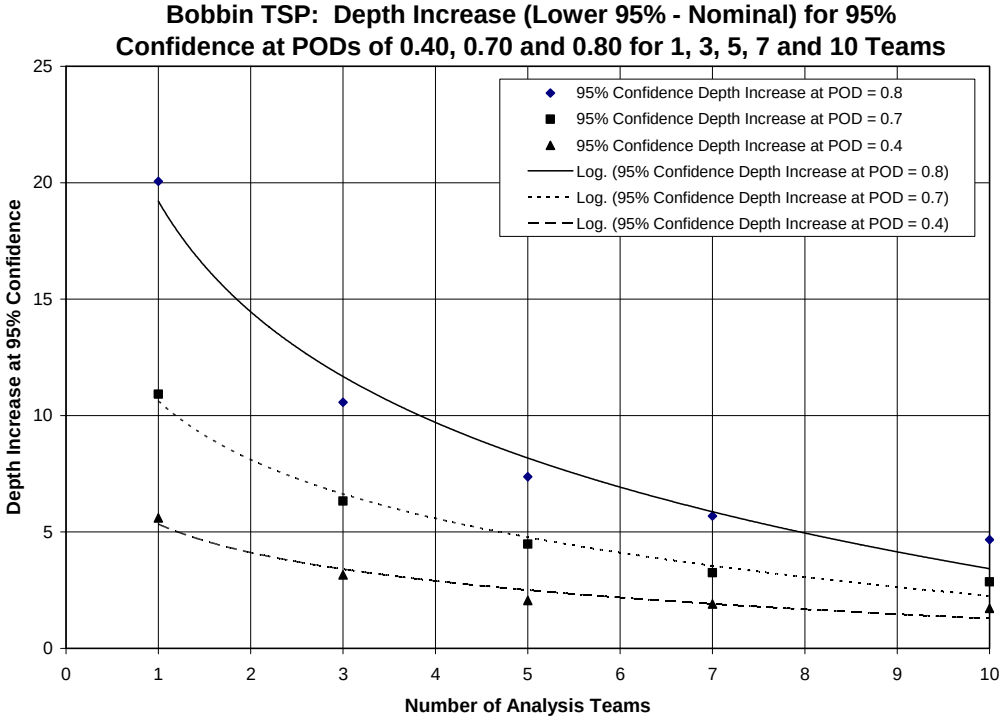

Figure 7-3
Bobbin Depth Increase (Lower 95% - Nominal) for Lower 95% Confidence at PODs of 0.40, 0.70 and 0.80

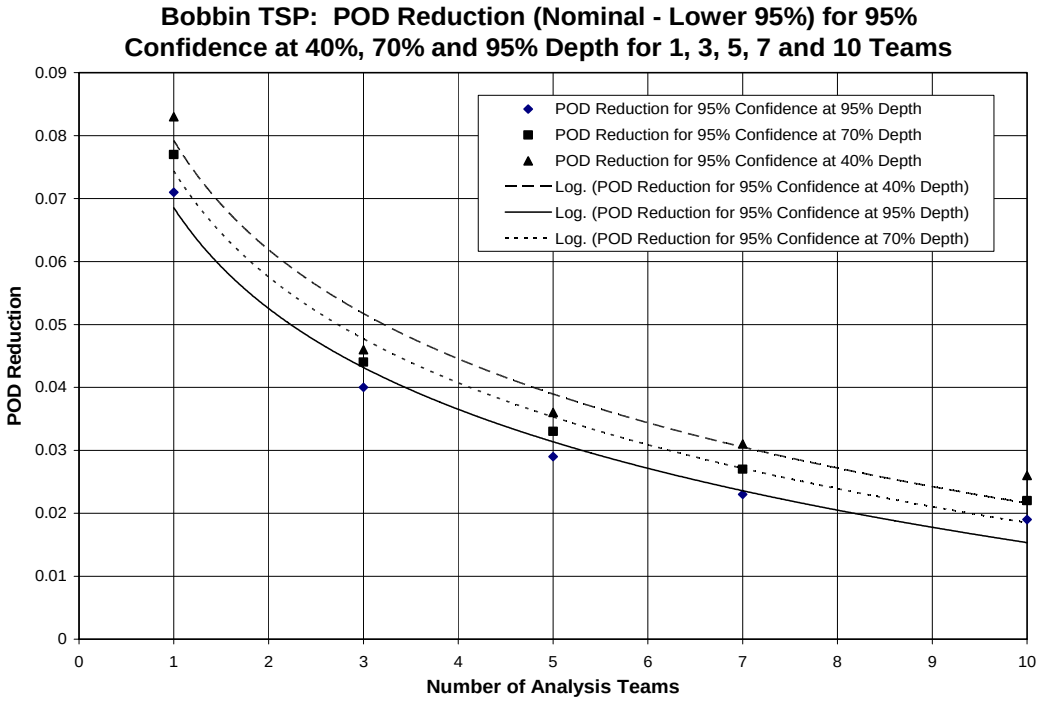

Figure 7-4
POD Reduction (Nominal - Lower 95%) for Lower 95% Confidence at Depths of 40%, 70% and 95%

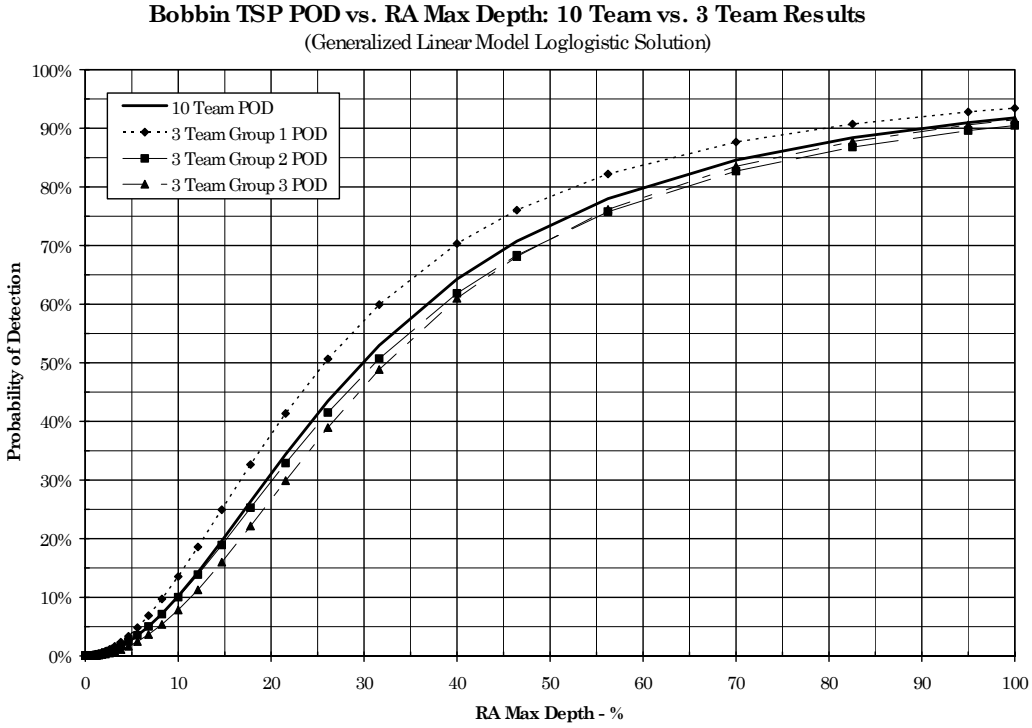

Figure 7-5
Bobbin TSP: Comparison of Nominal 10 Team POD with Three Team Sample PODs

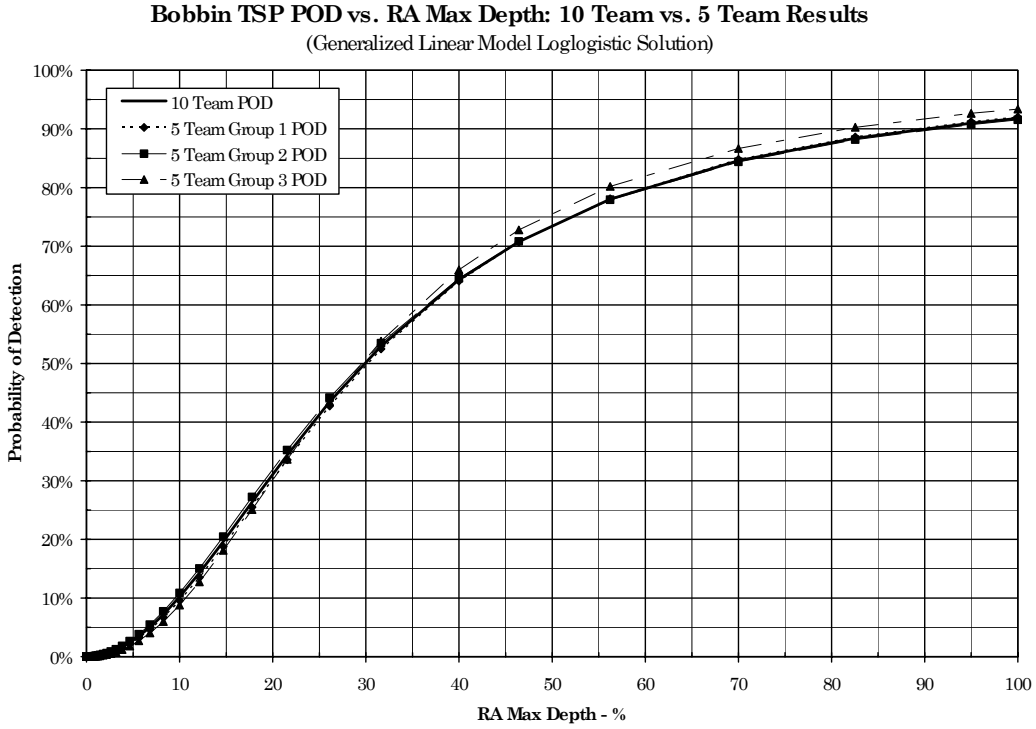

Figure 7-6
Bobbin TSP: Comparison of Nominal 10 Team POD with Five Team Sample PODs

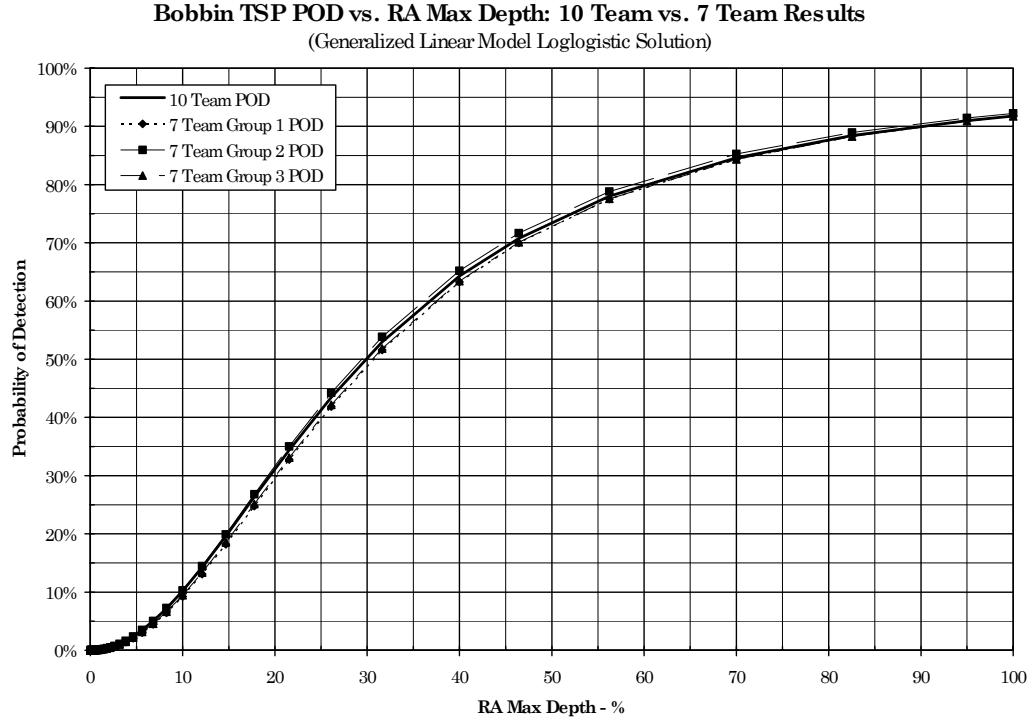

Figure 7-7
Bobbin TSP: Comparison of Nominal 10 Team POD with Seven Team Sample PODs

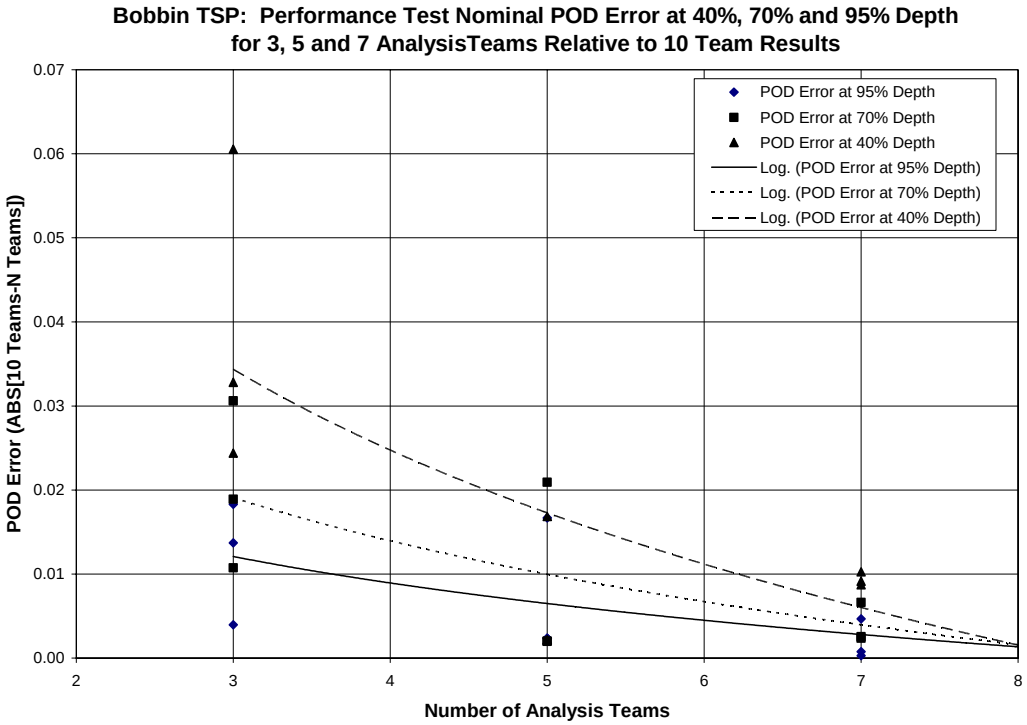

Figure 7-8
Nominal POD Error Relative to 10 Team Results at Depths of 40%, 70%, and 95%

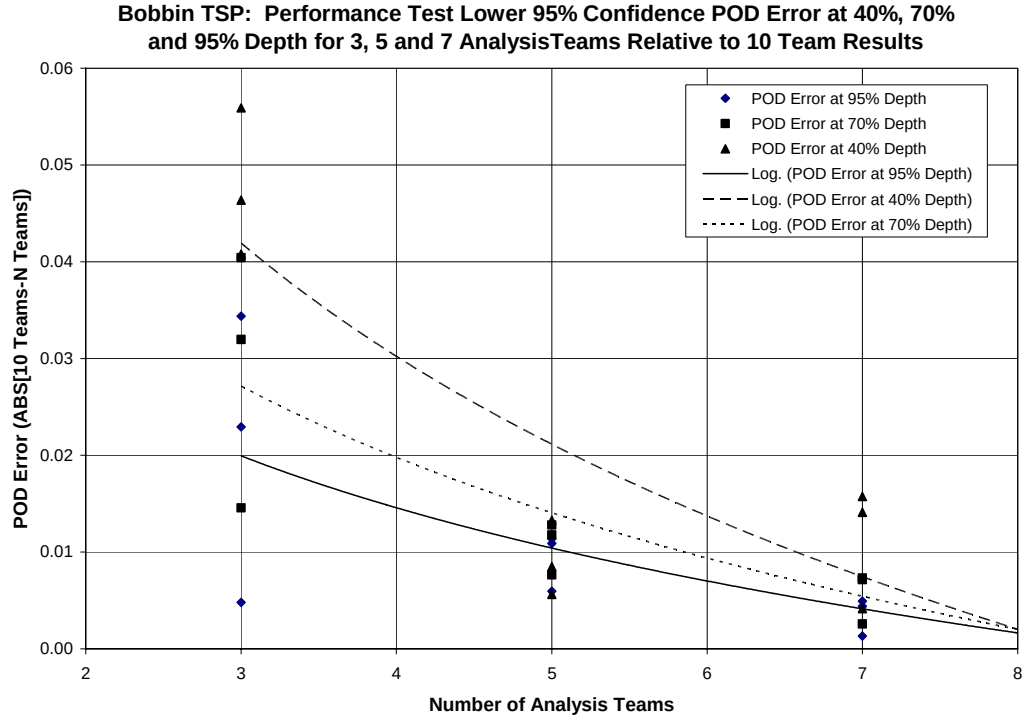

Figure 7-9
Lower 95% Confidence POD Error Relative to 10 Team Results at Depths of 40%, 70%, and 95%

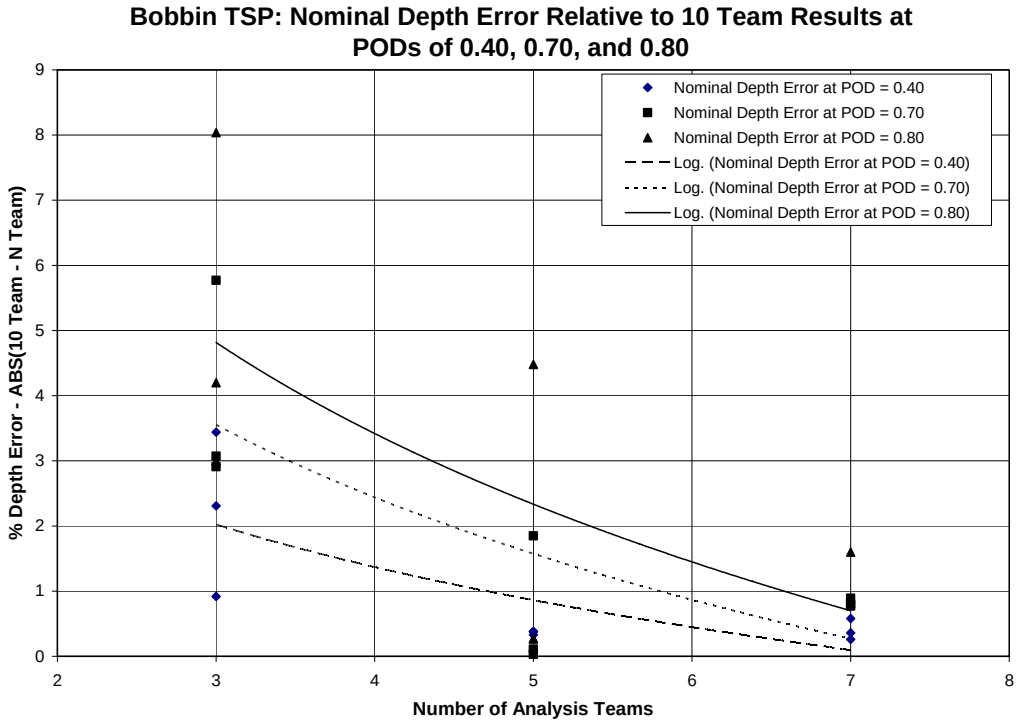

Figure 7-10
Nominal Depth Error Relative to 10 Team Results at PODs of 0.40, 0.70, and 0.80

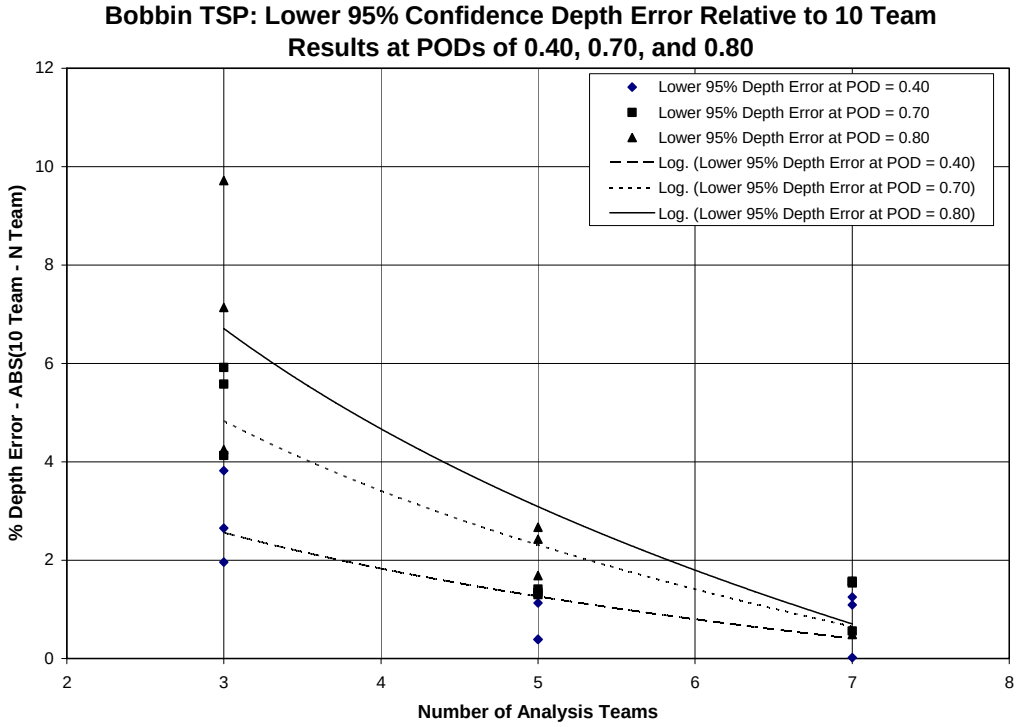

Figure 7-11
Lower 95% Depth Error Relative to 10 Team Results at PODs of 0.40, 0.70, and 0.80

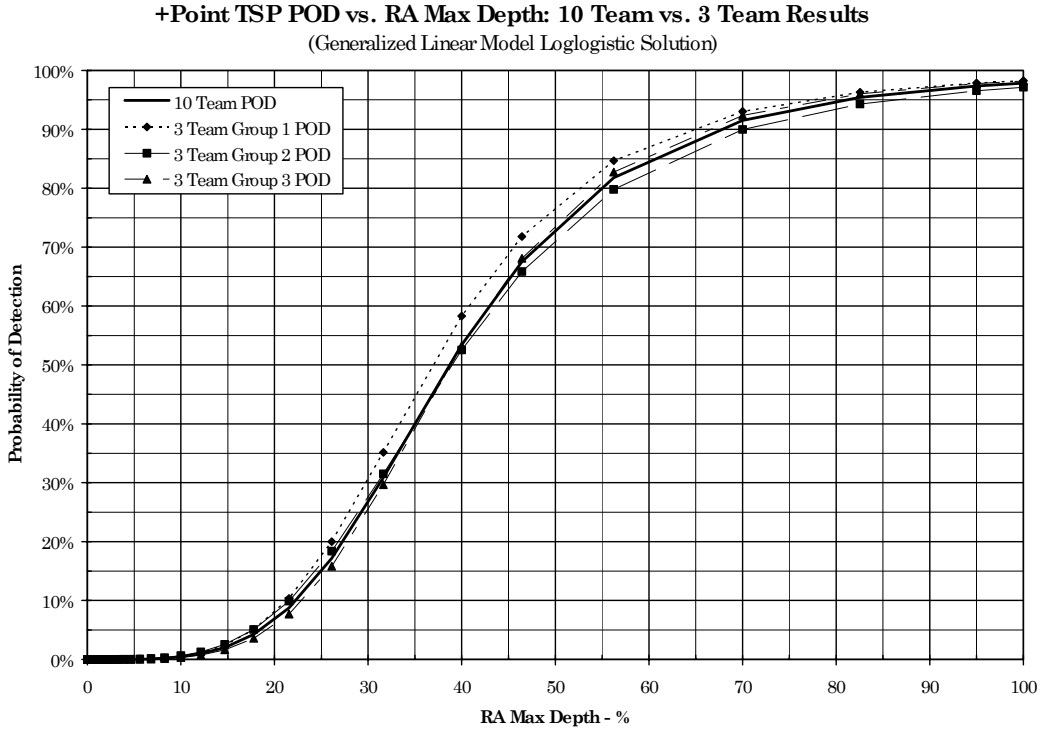

Figure 7-12
+Point TSP: Comparison of Nominal 10 Team POD with Three Team Sample PODs

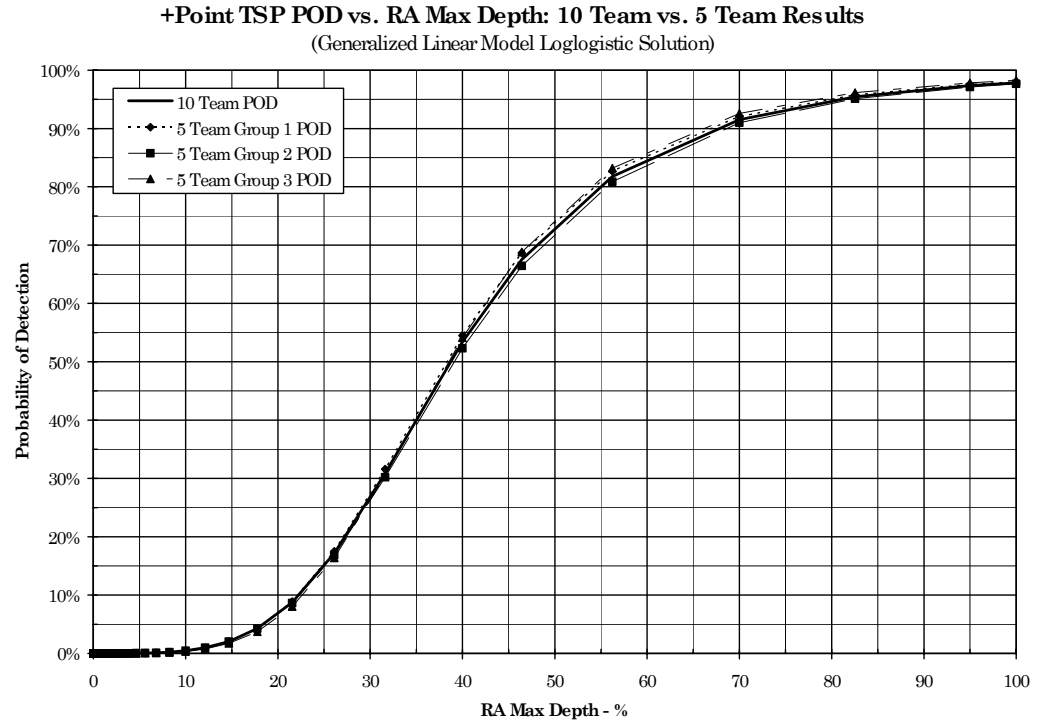

Figure 7-13
+Point TSP: Comparison of Nominal 10 Team POD with Five Team Sample PODs

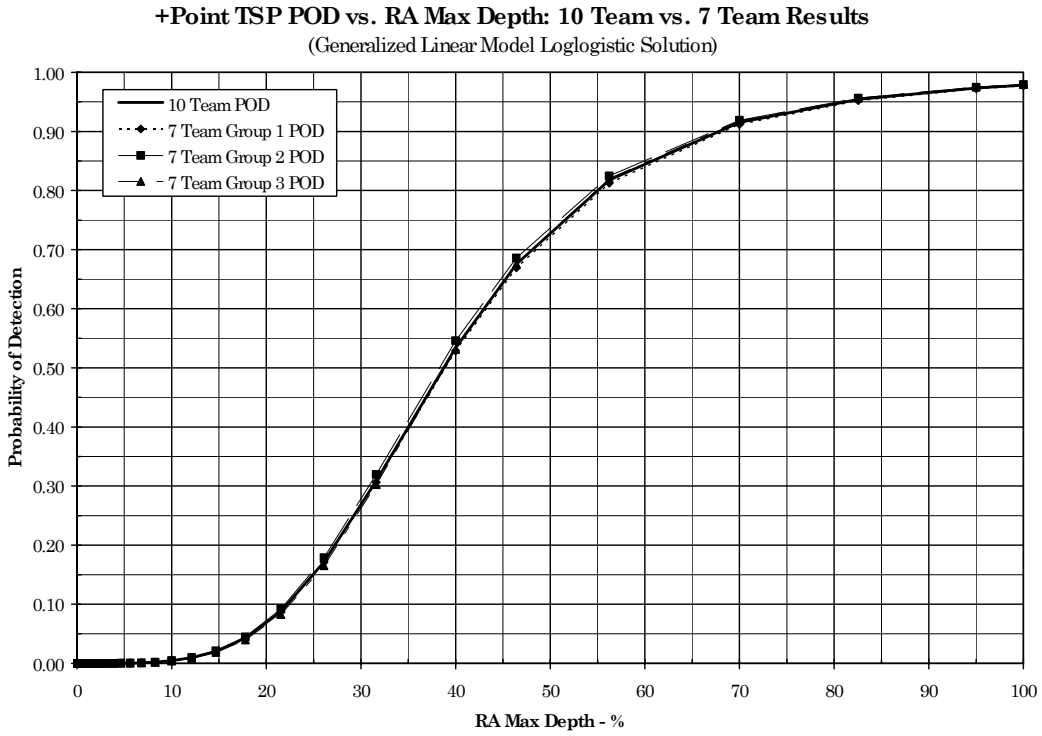

Figure 7-14
+Point TSP: Comparison of Nominal 10 Team POD with Seven Team Sample PODs

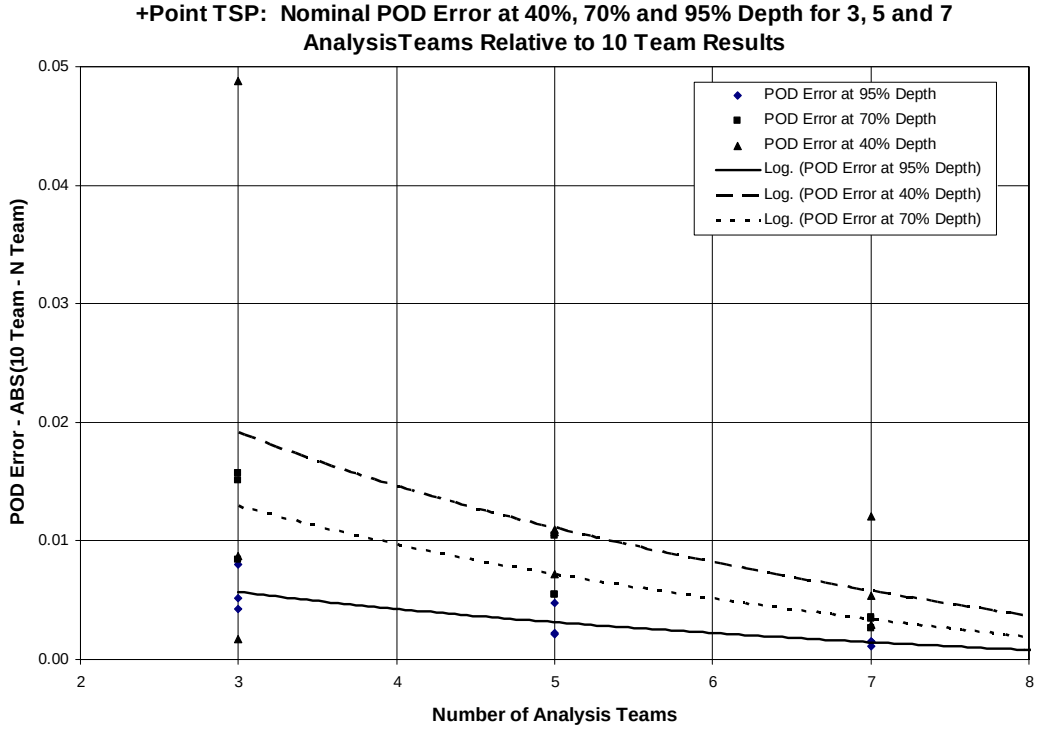

Figure 7-15
+Point TSP: Nominal POD Error Relative to 10 Team Results at Depths of 40%, 70%, and 95%

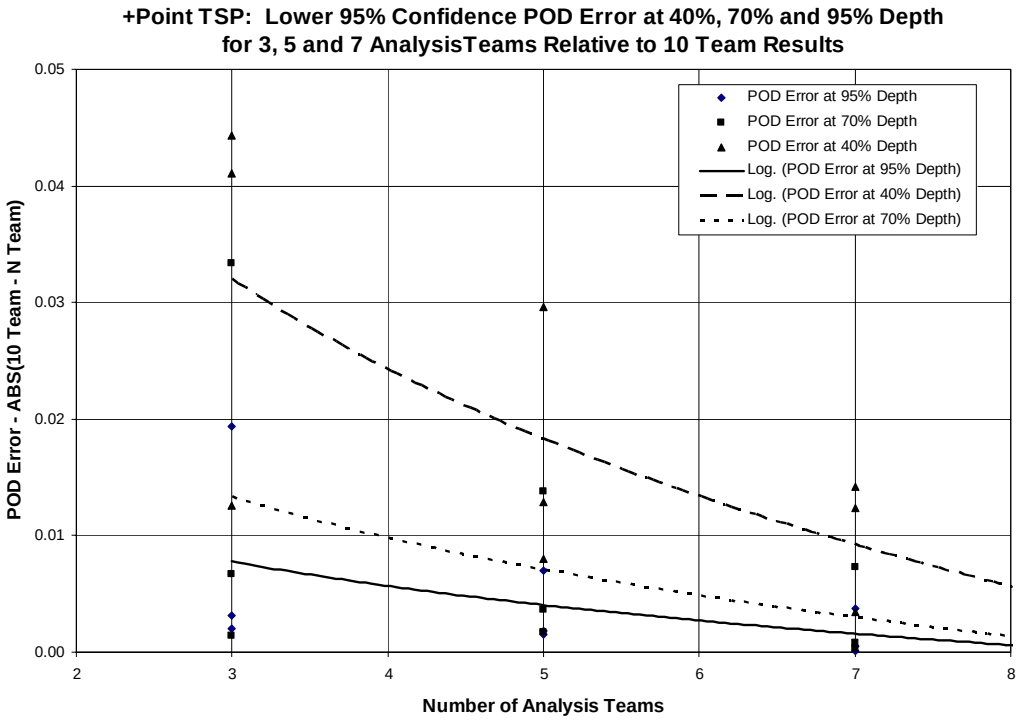

Figure 7-16
+Point TSP: Lower 95% Confidence POD Error Relative to 10 Team Results at Depths of 40%, 70%, and 95%

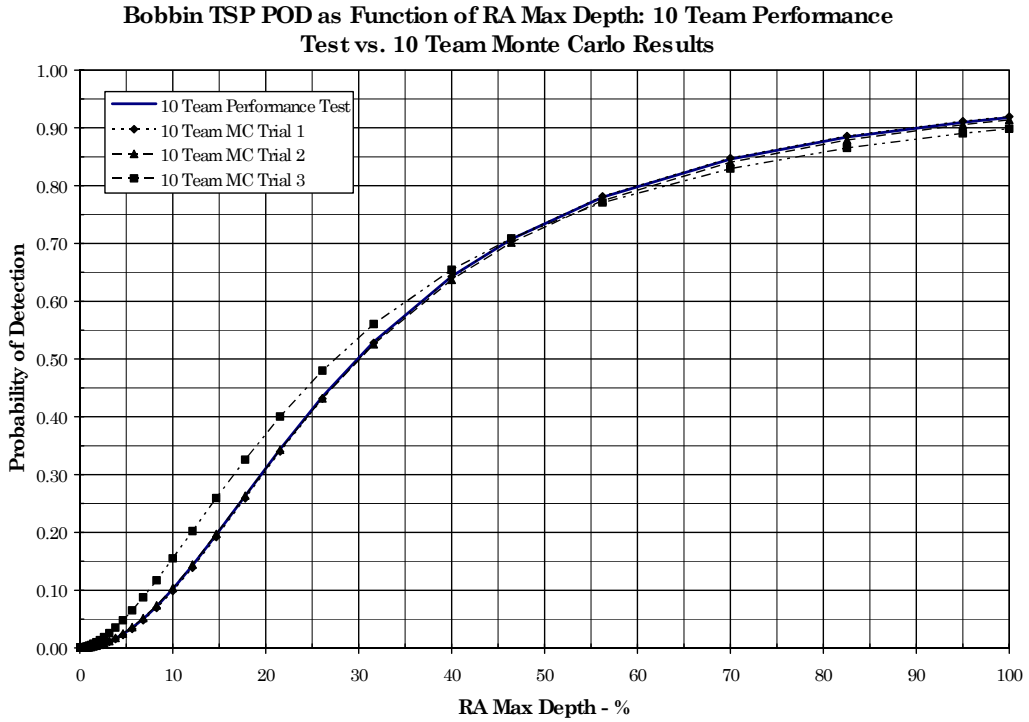

Figure 7-17
Bobbin TSP: Comparison of 10 Team Performance Test POD with 10 Team Monte Carlo Trials

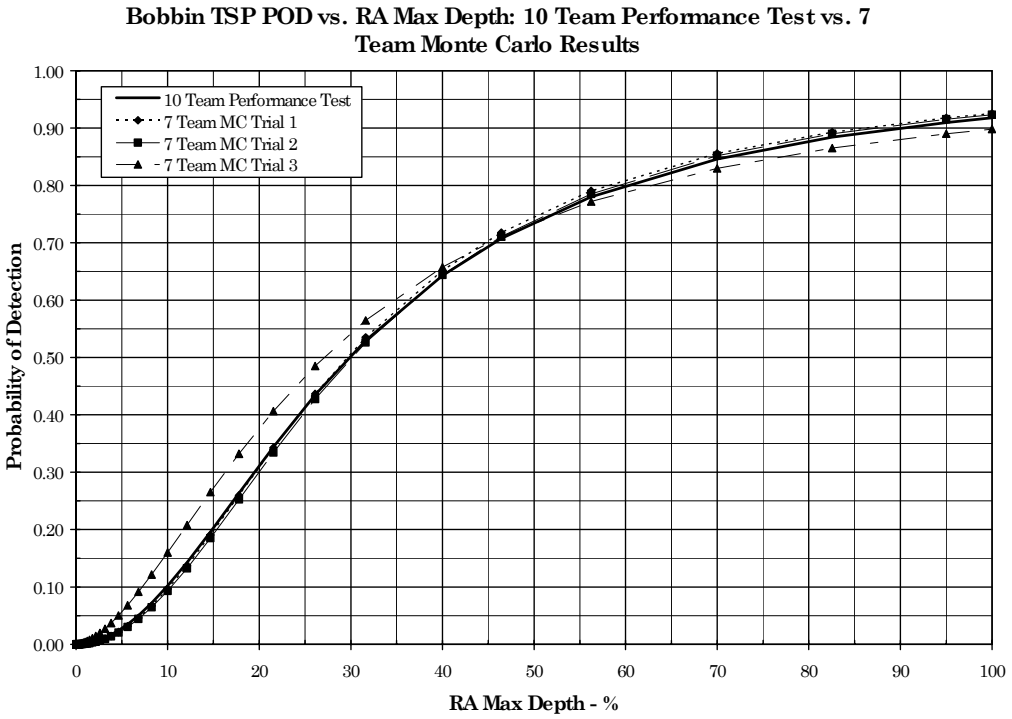

Figure 7-18
Bobbin TSP: Comparison of 10 Team Performance Test POD with 7 Team Monte Carlo Trials

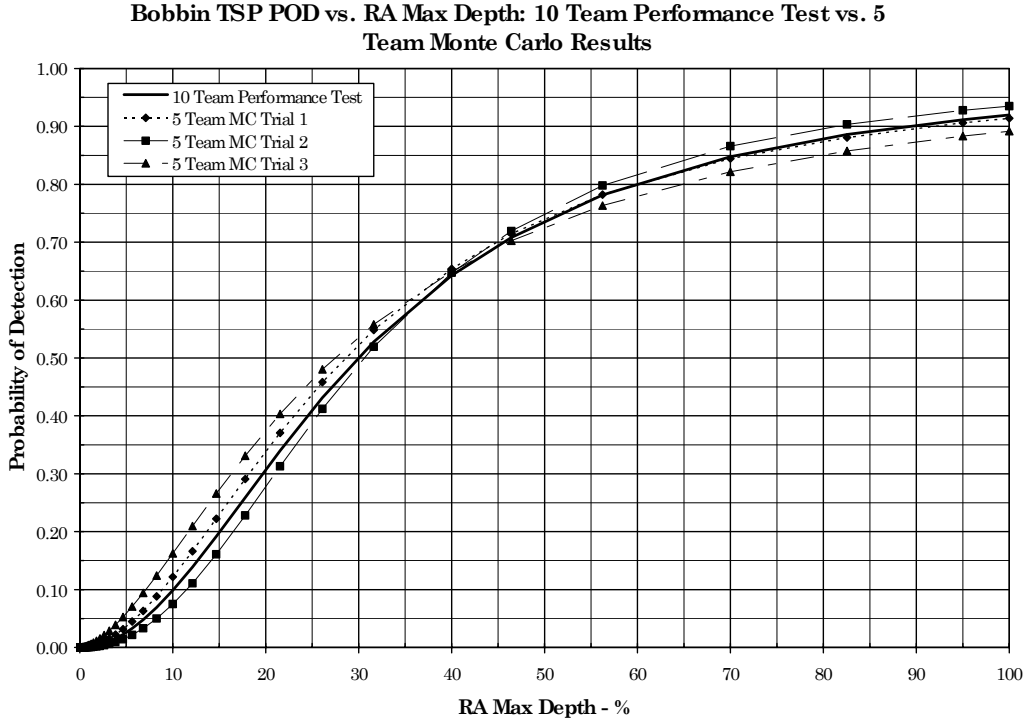

Figure 7-19
Bobbin TSP: Comparison of 10 Team Performance Test POD with 5 Team Monte Carlo Trials

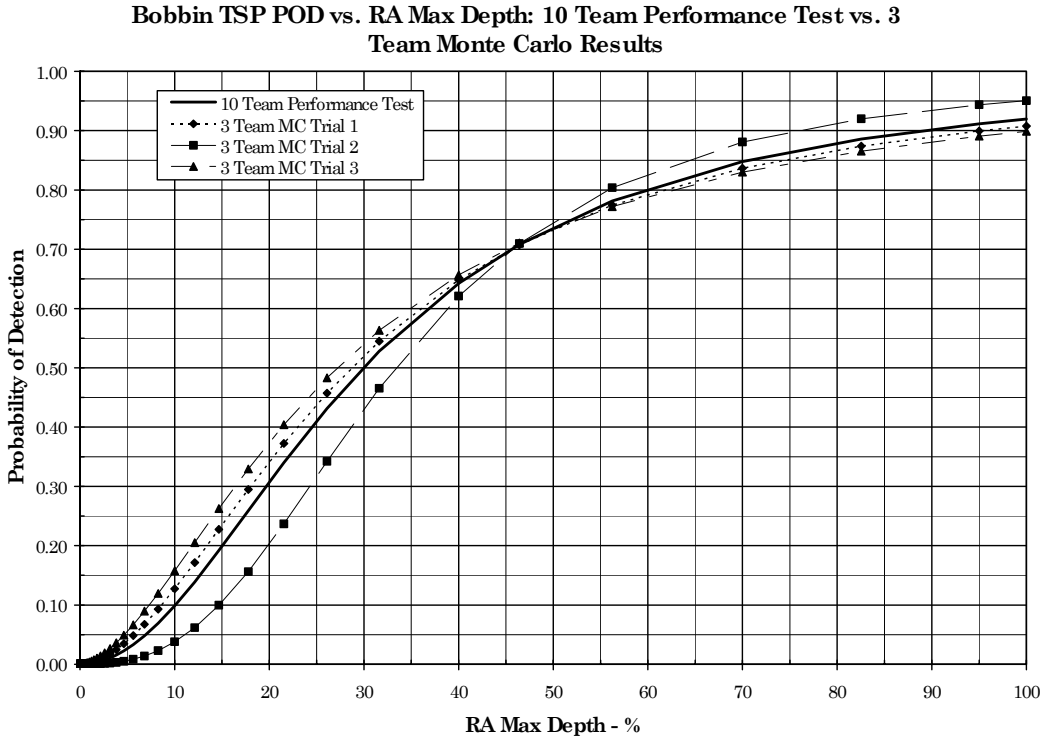

Figure 7-20
Bobbin TSP: Comparison of 10 Team Performance Test POD with 3 Team Monte Carlo Trials

8
REASSESSMENT OF NUMBER OF TEAMS REQUIRED FOR NDE SIZING TESTING

This section provides a reassessment of Section 6.0 for the number of NDE analysis teams required for sizing testing. The results of Section 6.0 (e.g., Figure 6-1) are based on 95% confidence limits for the multiplier applied to obtain the confidence bounds on the analyst standard deviation. Performance demonstration testing for sizing was conducted for 10 teams in 2005. However, the results showed excessive analyst variability, likely due to inadequate training of the NDE analysts, and the testing was repeated in 2006. Results of the repeat tests are not available at the time of this report revision.

8.1 General Considerations on Influence of Number of Analysis Teams on NDE Sizing Correlations

The unacceptable results of the 2005 NDE performance testing indicate that even 10 teams may not be adequate to compensate for outlier trends in the data when more than 1 or 2 team results show outlier trends. Consequently, the most important factor relating to the number of teams for sizing is to perform enhanced analyst training to improve consistency between analysts and reduce variability in the results. A reduction in the number of teams to less than 10 must be contingent on increased emphasis on analyst training.

Figure 8-1 provides a modification of Figure 6-1 based on normalizing the multiplier to 1.0 for 10 teams in order to more directly show the influence of reducing the number of teams below 10. The chi-squared term for the confidence multiplier on the standard deviation is defined as $[(n-2)/\chi^2]^{1/2}$ where n is the number of teams. For general applications, n is the number of data points so the influence of the number of teams is reduced when many flaws are included in the sizing testing. However, analyst variability dominates the spread about the sizing regression line and Figure 6-1 provides an estimate for the influence on the number of teams relative to the sizing uncertainty. The factor for 5 teams relative to 10 teams is 1.7.

When applying linear regression analysis for sizing correlations, a few outlier data points biased to either the high or low side of the regression line can have a significant influence on the nominal correlation as well as the uncertainty. The slope of the correlation can be strongly influenced by a few outliers at either the low or high end of the independent variable such as depth. Due to these considerations, adequate analyst training is more important than applying a large number (e.g., 10) of teams for the sizing tests.

8.2 NDE Sizing Sensitivity to Number of Analysis Teams Based on Monte Carlo Sampling

Monte Carlo analyses, similar to that applied for POD in Section 7.3, were performed to assess the general trend of nominal correlation errors and the standard deviation on the number of analyst teams. A reference sizing correlation, defined as "truth", with 100 data points and a standard deviation of 11% depth was applied for this analysis. Random sampling of the correlation for 100 data points was defined as the sizing results for one analysis team. The process was repeated 10 teams which were then combined to obtain 3, 5, 7 and 10 team results. Regression analyses were then applied to the grouped team sizing results to define a resulting nominal regression curve and standard deviation. This method was repeated three times to obtain 3 trials for each team group. It can be noted that this process leads to modest analyst variability since the sampling process does not lead to significant numbers of large outlier samples. The results would reasonably reflect that obtained in a performance test with well trained analysts such that analyst variability is modest.

Figure 8-2 shows the nominal sizing error at 40%, 70% and 100% depth as a function of the number of teams. Figure 8-3 shows the corresponding results for the error in the correlation standard deviation. The results indicate a negligible dependence of both the nominal error and standard deviation on the number of teams. These results support the intuitive expectation that when analyst variability is small, the performance test results would not be significantly dependent on the number of teams.

8.3 Recommendation on Number of Teams for NDE Sizing Testing

As noted above, adequate analyst training is expected to be more important for acceptable sizing correlations than testing based on a large number of teams. Even with adequate analyst training, it can be expected that some data points tending toward outlier results will be obtained. Five teams are recommended for NDE sizing testing to reduce the impact of a few outlier points on the mean correlation and associated uncertainties.

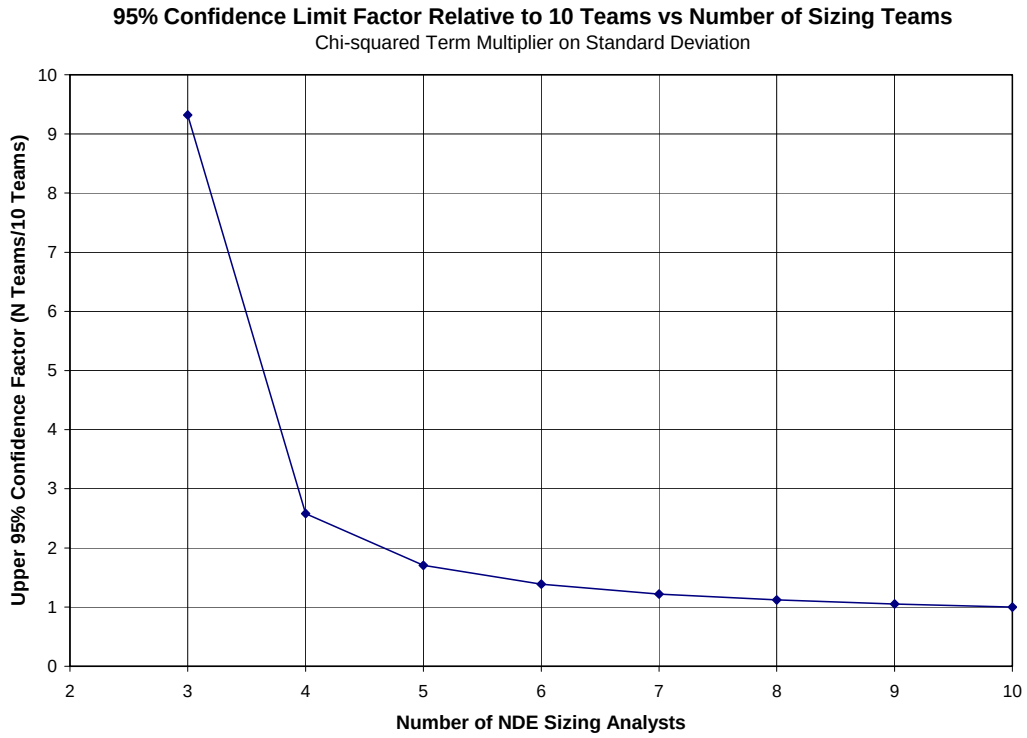

Figure 8-1
95% Confidence Factor on Sizing Uncertainty versus Number of Sizing Teams

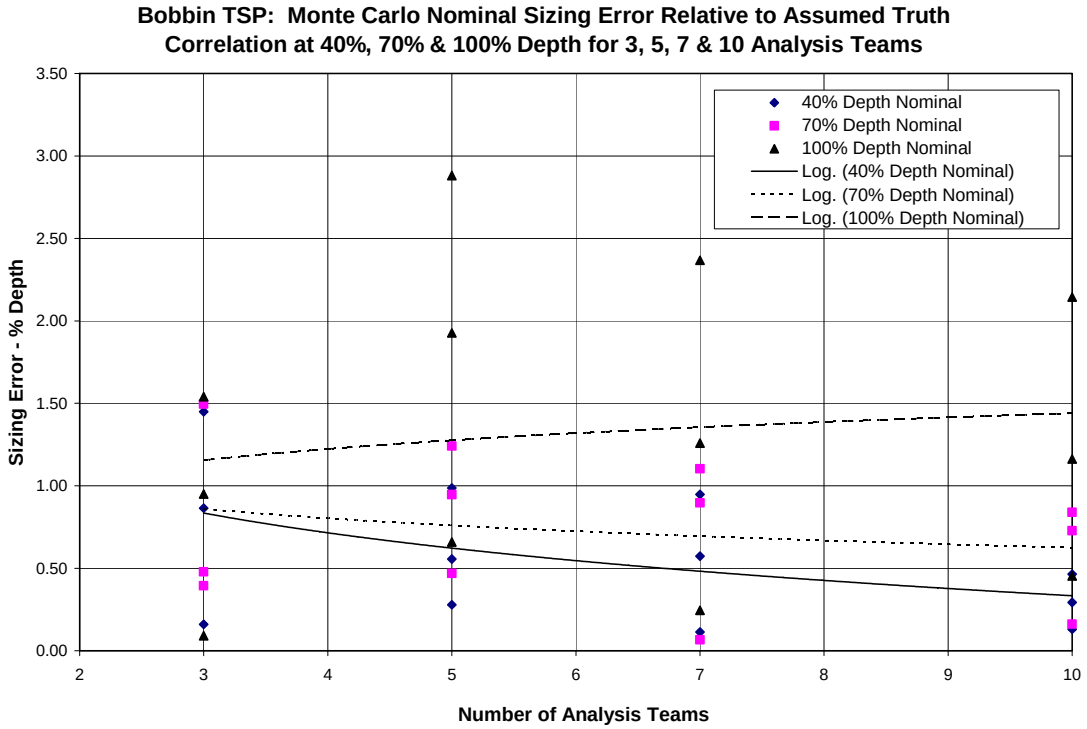

Figure 8-2
Bobbin TSP: Monte Carlo Nominal Sizing Error (Relative to Assumed Truth Correlation) at 40%, 70% and 100% Depth versus Number of Analysis Teams

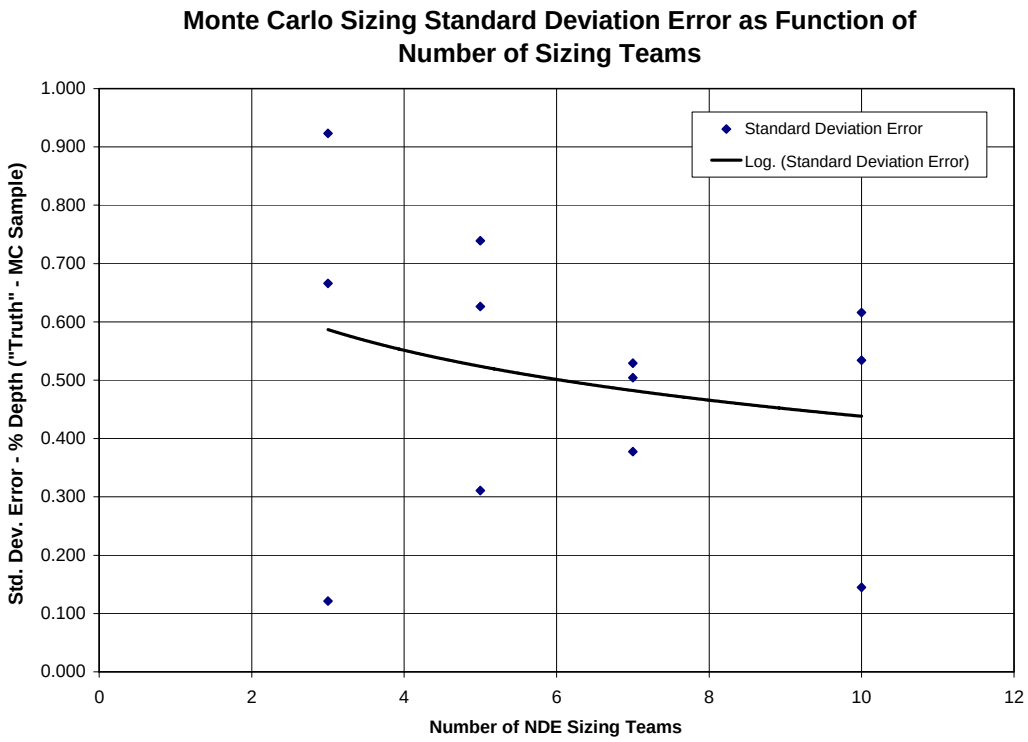

Figure 8-3
Bobbin TSP: Monte Carlo Sizing Standard Deviation Error (Relative to Assumed Truth Correlation) versus Number of Analysis Teams

9
SUMMARY AND CONCLUSIONS

The number of specimens required for an ETSS dataset to support POD development should be sufficient to control the sensitivity of the POD uncertainty to the number of specimens. The POD sensitivity to the number of specimens has been developed in Section 2 using Monte Carlo techniques. The results show a sharp knee in the dependence of the POD at a given depth on the number of flaws in the range of 30 to 40 specimens. Based on these results, the number of specimens for detection testing is recommended to be ≥ 40. If 40 specimens cannot be obtained for a given degradation mechanism and location, the ETSS must include a minimum of 20 specimens and the number of analyst teams should be increased beyond the minimum required as described below.

POD correlations are to be based on the results of multiple analyst testing simulating field inspections for which the call is based on the results of a resolution process (analysis team). As shown in Section 2, the effect of multiple resolution teams on inference of depth at a given POD can be significant for small numbers of teams. The beneficial effect appears to saturate beyond 10 to 15 resolution teams. The uncertainty in depth at a given POD (i.e., depth at POD = 0.95 at 95% confidence) is no longer a guideline for deterministic tube integrity analyses per Reference 4 and is not used in Monte Carlo analyses. The nominal POD value and the uncertainty in POD at a given depth are currently applied for tube integrity analyses. As shown in the reassessment in Section 7, the POD uncertainties at a given depth are significantly smaller and less dependent on the number of analyses teams than the depth at a given POD at a high POD value. The latter is due to the small slope for POD dependence on depth at high POD values. The results in the Section 7 reassessment of the sensitivity of POD on the number of analysis teams show that the minimum number of teams required for performance testing can be reduced from the Section 2 recommendation. The number of analysis teams required for an ETSS dataset to support POD development should be sufficient to control the sensitivity of the POD uncertainty to the number of specimens. There is an interaction on POD uncertainties between the number of analysis teams and the number of specimens. If the ETSS approaches the minimum number of flaws, it is recommended that the number of analyst teams be increased beyond the minimums described herein to help offset the increased uncertainty from the small number of flaws. The number of analysis teams for detection testing is recommended to be ≥ 5 for NDE performance testing.

As developed in Section 3, undetected flaw POD information is not to be indiscriminately added to the POD database. Due to the symmetry of functions such as log-logistic used for PODs, an excessive number of specimens at either the never-detected or always-detected ends of the data can affect the POD at the opposite end of the data (i.e., large numbers of undetected specimens at low depths can increase POD at larger depths). The number of indications below the accepted threshold of detection should not exceed 10% of the above threshold database for that mode of degradation. Similarly, to control the high POD curve, the number of detected data points at the high POD end above the highest depth non-detected indication in the database should not exceed

about 15% of the database. The numbers of specimens and analysis teams, as well as the distribution of specimens, for detection testing are summarized in Table 7-1.

Non-flawed or NDD specimens are to be included in performance testing for POD development to provide a constraint on the NDE analysts against overly conservative false call rates as described in Section 4. The NDD requirements are established to obtain a 90% confidence on an acceptable false call rate, which can be used to define an acceptable number of false calls for a given NDD population size. The population false call rate is estimated using a one-sided upper bound confidence limit for a binomial distribution. For bobbin coil analyses, conservative calls are to be encouraged to enhance the POD when rotating coils are to be used for flaw confirmation. RPC analyses should be less conservative than bobbin analyses since overcalls can lead to unnecessary tube repair. Since the primary and secondary analysts are encouraged to provide conservative calls, the false call requirements are applied to the results of the resolution analyst. If the false call rate for a resolution analyst's team exceeds the acceptance limits, the detection results for that team are not included in the POD evaluation for the associated dataset or ETSS. It is recommended that the false call rates for the primary and secondary analysts be evaluated for the performance test results although no acceptance criterion is applied for these results. The recommended false call criterion for the team results should be reviewed after the initial performance test results are evaluated to assess the adequacy of the criterion to avoid excessive conservatism in the testing. The required false call rates and number of required specimens as a function of the number of false calls are given in Table 7-2. The false call requirements are ≤20% at 90% confidence for bobbin detection and ≤10% at 90% confidence for RPC (rotating coils – pancake, +Point, etc.). Dependent on the difficulties in obtaining NDD specimens, the number of specimens can be increased by the increments given in Table 7-2 or larger increments to provide increased allowances for false calls. The minimum number of NDD specimens with no false calls permitted would be 11 specimens for bobbin detection and 22 specimens for RPC detection. An allowance for at least one false call is suggested but not required.

Based on the requirements developed in Section 5, Table 7-3 summarizes the requirements on the number and depth distribution for the samples to be used for performance testing to support NDE sizing correlation development. Separate requirements are given for sizing correlations applied for tube repair limits and for applications to tube integrity analyses only. Variations of a few percent in the depth ranges of Table 7-3 are acceptable. For correlations with variables other than depth, such as length or crack area, the variable should be assessed over approximately 1/3 spans of the data with the intent of obtaining a distribution with the minimums of Table 7-3 over each of the three spans.

The issue of concern in terms of the number of required analysts for NDE sizing is the analyst variability component. A sufficient number of analysts are required to obtain an estimate of analyst variability that can be expressed as a standard deviation as described in Section 6. The confidence bound decreases rapidly prior to a sample size of 10 to15 analysts. The results apply to the number of NDE sizing teams, which typically include a sizing analyst and an independent technical reviewer (ITR). For 5 teams, the 95% confidence factor on the standard deviation can be up to a factor of 1.7 when a small number of specimens are included in the testing. For 3 teams, this factor could be as high as 9 so that 5 teams represents a large reduction in the sizing uncertainty at 95% confidence relative to fewer teams. The revised number of analysis teams for sizing testing is recommended to be ≥ 5. However, when limited to 5 teams, it is important that

adequate training of NDE sizing analysts be performed to improve consistency of the sizing results and reduce variability in the analysis results (i.e., reduce potential for outlier analyses). A minimum of 5 teams is recommended to reduce the impact on the sizing correlations of any remaining outlier behavior after enhanced analyst training.

Table 9-1
Required Detection Distribution and Number of Analyst Teams for Performance Testing to Develop POD Correlations

Number of Specimens	
Minimum	≥ 20
Preferred	≥ 40
Detection Range	
Undetected	≤ 10% of Specimens
Detected Above Largest Undetected	≤ 15% of Specimens
Number of Analysis Teams	≥ 5

Table 9-2
Required False Call Rates and Number of NDD Specimens for 90% Confidence on False Call Rate

Probe	Acceptable Resolution Analyst False Call Rate	Acceptable Number of False Calls	Required Number of NDD Specimens
Bobbin	≤20%	0	11
		1	18
		2	25
		3	32
RPC	≤10%	0	22
		1	38
		2	52
		3	65

Table 9-3
Required Number and Maximum Depth Distribution of Samples for Performance Testing to Develop NDE Sizing Correlations

	Application for NDE Sizing Correlation	
Maximum Depth Range	**Sizing Supporting Tube Repair Requirements**	**Sizing Supporting Tube Integrity Analyses Only**
0 – 35%	≥ 10	≥ 7
36% - 65%	≥ 10	≥ 6
66% - 100%	≥ 10	≥ 7
Total	≥ 30	≥ 20
Number of Analysis Teams	≥ 5(1)	≥ 5(1)

Notes: 1. Training of NDE sizing analysts is particularly important, when limiting number of teams to 5 teams, in order to improve consistency between analysts and reduce variability in sizing results.

10
REFERENCES

1. Pressurized Water Reactor Steam Generator Examination Guidelines: Rev. 6, Requirements. EPRI, Palo Alto, CA: 2002. 1003138.
2. Berens A.P, "NDE Reliability Data Analysis," ASM Metals Handbook Volume 17 Nondestructive Evaluation and Quality Control, 9th Ed., American Society of Metals, 1989.
3. Nondestructive Evaluation System Reliability Assessment. Department of Defense Handbook MIL-HDBK-1823, April 39, 1999.
4. EPRI Report 1012987, Revision 2, "Steam Generator Integrity Assessment Guidelines", Final Draft Report, July 2006

Export Control Restrictions

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

The Electric Power Research Institute (EPRI), with major locations in Palo Alto, California, and Charlotte, North Carolina, was established in 1973 as an independent, nonprofit center for public interest energy and environmental research. EPRI brings together members, participants, the Institute's scientists and engineers, and other leading experts to work collaboratively on solutions to the challenges of electric power. These solutions span nearly every area of electricity generation, delivery, and use, including health, safety, and environment. EPRI's members represent over 90% of the electricity generated in the United States. International participation represents nearly 15% of EPRI's total research, development, and demonstration program.

Together...Shaping the Future of Electricity

Program:

Nuclear Power

Printed on recycled paper in the United States of America

1014756

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com