NASA/TP–2011-217176

Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications

Edward R. Generazio
Langley Research Center, Hampton, Virginia

September 2011

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at *http://www.sti.nasa.gov*

- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at 443-757-5803

- Phone the NASA STI Help Desk at 443-757-5802

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7115 Standard Drive
  Hanover, MD 21076-1320

NASA/TP–2011-217176

Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications

Edward R. Generazio
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

September 2011

Acknowledgments

The author wishes to thank Ward D. Rummel of D & W Enterprises for his unwavering support, encouragement, recommendations, and testing of DOEPOD software, and Dr. William Q. Meeker of Iowa State University for reviewing the manuscript, clarifying statistical concepts and procedures, and providing guidance on Monte Carlo testing.

Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320
443-757-5802

Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications

E. R. Generazio[1]

[1]National Aeronautics and Space Administration, Hampton, VA 23681

ABSTRACT. The capability of an inspection system is established by applications of various methodologies to determine the probability of detection (POD). One accepted metric of an adequate inspection system is that for a minimum flaw size and all greater flaw sizes, there is 0.90 probability of detection with 95% confidence (90/95 POD). Directed design of experiments for probability of detection (DOEPOD) has been developed to provide an efficient and accurate methodology that yields estimates of POD and confidence bounds for both Hit-Miss or signal amplitude testing, where signal amplitudes are reduced to Hit-Miss by using a signal threshold Directed DOEPOD uses a nonparametric approach for the analysis or inspection data that does require any assumptions about the particular functional form of a POD function. The DOEPOD procedure identifies, for a given sample set whether or not the minimum requirement of 0.90 probability of detection with 95% confidence is demonstrated for a minimum flaw size and for all greater flaw sizes (90/95 POD). The DOEPOD procedures are sequentially executed in order to minimize the number of samples needed to demonstrate that there is a 90/95 POD lower confidence bound at a given flaw size and that the POD is monotonic for flaw sizes exceeding that 90/95 POD flaw size. The conservativeness of the DOEPOD methodology results is discussed. Validated guidelines for binomial estimation of POD for fracture critical inspection are established.

INTRODUCTION

Recently it was reported[1, 2, 3, 4] that Design of Experiments for Determining Probability of Detection Capability (DOEPOD) methodology provided a unique perspective on understanding probability of detection data. The DOEPOD methodology is based on the use of a nonparametric binomial statistical model. It was reported that the inspection data can be categorized into a series of numbered classes, depending on the structure of the data. Classes CASE 1 and CASE 1+ are classes with data that exhibit point estimates that are monotonically increasing with flaw size are identified. Classes CASE 1#, CASE 1*, CASE 2, CASE 4, CASE 5, CASE 6, and CASE 7 are data that either exhibits non-monotonic point estimates with flaw size or there is insufficient data to make a determination of monotonicity. The identification of different cases of inspection data allows development of an intuitive understanding that provides guidance on qualifying nondestructive inspection technologies. One accepted metric of an adequate inspection system is that for a minimum flaw size and all greater flaw sizes, there is 0.90 or greater probability of detection with 95% confidence (90/95 POD). There is often an assumption that the POD is increasing with flaw size. This assumption, however, is not always justified and the DOEPOD methodology does not require this assumption. The origin of using 90/95 POD as a metric for inspection capability is found in Mil-HDBK-5H[5] where the 90/95 bound (T90) for acceptable B-basis material properties is defined. T90 is the "Statistically based lower tolerance bound for mechanical property such that at least 90 percent of the population is expected to exceed

T90 with 95 percent confidence."[5] There are other more precise definitions of confidence intervals, clarifying that 95% confidence is a property of the procedure for constructing a statistical interval, and not to the observed interval itself[6].

It is important to define the difference between verification and validation. Verification is a demonstration that the process and procedures meet the specified requirements. Coding the physics correctly is an example for computer modeling verification. Validation is a demonstration that the process and procedures yield specific quantitative results within the specified requirements. Demonstrating that the physics is correct is an example for computer modeling validation.

The DOEPOD binary data-based methodology is designed to yield two main results. The first result to identify whether or not the data provide a binomial-based 95% lower confidence bound that is at least 0.90 at a flaw size, known as the 90/95 flaw size. The second result is to identify whether or not the POD is monotonically increasing for flaw sizes greater than the 90/95 POD flaw size.

A validation demonstration of the DOEPOD methodology is needed and a Monte Carlo simulation study provides a basis for an acceptable demonstration. The simulation study is to show that, given a set of inspection data, the process and procedures used in DOEPOD will, with high probability identify a 90/95 point that is at least as big as the true flaw size giving a POD of 0.90 and to correctly determine if the POD is monotonically increasing for flaw sizes greater than the 90/95 POD flaw size.

Binomial-based statistical analyses[1, 2, 3, 4] of sample sets may indicate that a 90/95 point exists at a particular flaw size. However, this does not assure that 0.90 POD with 95% confidence also exists for larger flaw sizes. For fracture-critical applications, it needs also to be demonstrated that POD is at least 0.90 for a range of flaws sizes that is larger than the identified 90/95 point. This extension may be made when appropriate data is included. Because the statement, "POD is at least 0.90 with 95% confidence," generally refers to the confidence bound at only one flaw size, it is important to define precisely the 90/95 POD acronym. For following work and the analyses executed in DOEPOD, the acronym 90/95 POD refers to, "By this procedure, there is a minimum flaw size, X_{pod}, at which POD meets or exceeds 0.90 with 95% confidence, and that POD also meets or exceeds 0.90 with 95% confidence for all flaw sizes larger than X_{pod}."

Validating that POD exceeds 0.90 with 95% confidence for all flaw sizes larger than the identified minimum flaw size, X_{pod}, is of fundamental concern. This is of particularly importance where the POD at larger flaws sizes may be constant, oscillating, or even decreasing as a function of flaw size due to the physics of the inspection. The fundamental, issue is the lack of *a priori* knowledge about the shape of the POD function, making it impossible to specify, *a priori*, an economical investigation plan specifying the number, distribution, and size range of flaws needed to assure that POD meets or exceeds 0.90 with 95% confidence over a given range of interest. These sample requirement issues are not unique to binomial-based point estimates. This also seen more generally in parametric-model based POD methods[7, 8] where *a priori* knowledge about the shape of the POD function is used to recommend sample requirements. Since these sample requirements are of critical practical importance, the need for a procedure to address sample requirements is needed.

The motivation for developing the DOEPOD methodology stems from the author's prior attempts to verify the accuracy of POD curves[7, 8, 9], developed from parametric-model based POD methods, by utilizing binomial–based probability models that do not require a monotinicity assumption to estimate POD. It was noticed that binomial point estimates of POD often varied dramatically below the 90/95 point [7, 9] provided by the commonly-used two parameter logistic regression model. These variations lead to a concern that if binomial point estimates of POD are varying dramatically below the 95% lower confidence bound, then there could be some doubt that POD is at least 0.90 at that flaw size.

Since the actual form of POD functions is unknown and varies with each inspection system and application, the author used existing inspection data sets as an initial basis to explore the large flaw sample requirements. A statistical "Delete-M" jackknife[10] approach is used to generate subsamples from the existing data sets that have various predetermined numbers of randomly selected large flaws. Basically, the question being addressed by the jackknife approach is what if only 2, or 3, … or 30 large flaws are included in the analysis. An alternative approach that may be pursued to optimize the large flaw requirements is to use simulation data sets based on a given analytical POD function. The optimization using analytical POD functions is not reported here, and is future work to be explored. A Monte Carlo evaluation of existing data sets is used to identify the minimum sample requirements, and includes the use of a simple binomial point estimate statistical model of POD, generation of random jackknife subsamples data sets, subsequent DOEPOD analysis of the "Delete-M" jackknife subsamples, and aggregation and analysis of results.

The jackknife method used here may also be considered to be a repeated "Delete-by-One" jackknife, where the effect of sample size (i.e., the number of large flaws) is explored by systematically varying the numbers of samples available. This effectively generates jackknifed sample subsets of varying sample sizes. The DOEPOD analysis includes requirements that are of particular importance to practical applications for fracture critical inspections. Specifically,

1) If, for a narrow range of flaw sizes, there is a binomial-based lower 95% confidence bound on POD that exceeds 0.90, then the largest flaw in this range is identified as the minimum flaw size, X_{pod}.

2) There must be a binomial-based lower 95% confidence bound on POD that exceeds 0.90 for groups of flaws having overlapping size intervals and also having flaw sizes within the range X_{pod} to at least $3xX_{pod}$. $3xX_{pod}$ is selected here as a representative bound that reflects current typical data sets[3].

3) There must not be any misses at large flaws.

The first requirement assures that there is a grouping of flaws that is similar in size and detectability, as required for binomial statistics.[1, 2, 3, 4] The second requirement assures that for a sufficient range of large flaws, POD meets or exceeds 0.90 with 95% confidence for overlapping groups of large flaws. This does not imply that the POD is monotonically increasing for large flaws, rather only that there is a binomial-based lower 95% confidence bound on POD which exceeds 0.90 over the range of the overlapping groups. It should be remembered that there are physical reasons why the POD may

exhibit oscillations with flaw size, e.g., near and far field ultrasonic and eddy current footprint effects. Therefore, the DOEPOD methodology does not assume that the POD is increasing, but rather checks to see if the POD can be demonstrated to meet or exceed 0.90 with 95% confidence for flaw sizes greater than X_{pod}. The third requirement adds a conservative constraint that flaws with sizes greater than the minimum flaws size X_{pod} must not be missed. Operationally, it is possible that such flaws will be missed in practical NDE capability testing for fracture critical applications. If flaws having a flaw sizes greater than X_{pod} are missed, then all such misses are flagged as exceptions that need to be explained. If the number of hits is large enough to compensate for any misses that might be detected (i.e., the resulting 95% lower confidence bound is at least 0.90), then the DOPEPOD methodology will indicate that the X_{pod} flaw size as acceptable, while flagging that the misses should be evaluated to determine if there are other physical reason, such as unclean samples, or procedurally error, etc. When qualifying inspectors, other experience based conditions are placed on the inspection results. The DOEPOD methodology does not support inspector qualification when large flaws are missed. As a result the DOEPOD methodology, X_{pod} will not exceed the size of the largest flaw missed.

The DOEPOD methodology is based on binomial statistics and provides point estimates of POD and companion lower confidence bounds on POD. No continuous functional form of POD versus flaw size is provided by the DOEPOD analysis. A comparison of DOEPOD analysis results with another popular curve fitting POD method[7, 9] is instructive and included in this work. There are a variety of statistical models and confidence bound procedures available, and for this comparison the estimated two parameter binary logistic regression models and corresponding lower confidence bounds compared are detailed in references 4 and 6. Point estimates of the logistic model POD may be obtained by the method of maximum likelihood estimation[11, 12].

The acronym Logit-ML is defined here and used throughout the following to denote the maximum likelihood estimation of the two parameter Logit statistical model[7, 9] and companion confidence bound procedures[7, 9].

A DOEPOD analysis on 437 POD data sets[7] was performed to verify the conservativeness of DOEPOD 90/95 POD values, relative to the Logit-ML method[7, 9]. There is no *a priori* reason why a nonparametric method should be conservative, relative to fitting a parametric model. However, there are three aspects to this issue that result in the DOEPOD methodology providing conservative POD results. First, it is preferred to explore nonparametric methods initially as they require minimal assumptions and thus resulting conclusions are based on a more solid foundation[6]. Second, it may not be possible to validate the assumptions behind parametric methods of estimating POD that are used to quantify the capability of inspection. Lack of validation can result in an implied validation, yielding results that may be incorrectly optimistic[13]. Third, by design and for specific applications to fracture critical inspections, the DOEPOD methodology includes requirements that are expected to yield conservative results. When considering these three aspects, it is expected that the results obtained by use of the DOEPOD methodology are conservative with respect the results of obtained by the Logit-ML method[7, 9].

BACKGROUND

The DOEPOD methodology utilizes the concept of "point estimate Probability of a Hit" (POH) at a given flaw size. Using the binomial distribution model, one can use the number of detections out of a certain number of inspections to compute a point estimate and a lower confidence bound on POD. Prior work[14, 15] used a selection of arrangements for grouping flaws of similar characteristics. Yee[14] used smoothing optimized probability and overlapping sixty point methods, grouped by number of flaws into a class and by cumulative sums of fixed flaw size class intervals, while Rummel[15] used fixed class widths. These binomial statistical methods have lead to the acceptance of using the 29 out of 29 (29/29) binomial confidence bound [14, 15] method, in combination with validation that the POD is increasing with flaw size, to meet the requirements of MSFC-STD-1249[16] and NASA-STD-5009[17]. The National Aeronautics and Space Administration has successfully used the (29/29) binomial-based rule for all fracture critical components of the Space Transportation System (Space Shuttle), International Space Station, and launch systems since the year 1970. The DOEPOD methodology extends previous work using the binomial distribution for estimating POD capability by adding the concept of maximizing the lower confidence bound as the driver for validating that 90/95 POD has been demonstrated at a minimum flaw size. When DOEPOD indicates that 90/95 POD is demonstrated at a minimum flaw size and for all larger flaw sizes, this satisfies the requirement for critical applications where validation of inspection systems, individual procedures, and qualification of operators are required. A DOEPOD analysis is useful even when a full POD curve[7] is estimated and is without formal internal and external validation [13]. It was noted in prior work[1, 2, 3, 4] that the combined statistical procedures of DOEPOD required further investigation by Monte Carlo simulation to obtain a clear picture of the statistical properties of the procedures. This work attempts to support that validation.

DETERMINATION of X_{pod}

The determination of X_{pod} is described in prior work[1, 2, 3, 4] and is briefly described here. The DOEPOD methodology is based on the application of the binomial distribution to a set of flaws that have been grouped into size classes, where each class has a width. The classes are allowed to vary in width and start at 0.001 inches and increase in width by 0.001 inch increments. Classes start at the largest flaw and move toward the smallest flaw. Class length is used here to represent the flaws features of interest to allow for flaw depth, shape, volume, etc., to be used as the inspection criteria. The first class width group is assigned to the largest flaw in the data set. The largest flaw in any class width group is assigned as the identifier of the group. The DOEPOD methodology computes the binomial-based point estimate, which is called Probability of Hit (POH) and the corresponding 95% lower confidence bound (LCL) from the flaw data within class width group and conservatively associates it with the largest flaw in the group. The next moving class width group is determined by decrementing the upper and lower class lengths bounding the class width group by 0.001 inch. In this manner the class of uniform width is moved. The DOEPOD analysis again evaluates the POH and LCL obtained from the flaw data within this new class width group. This process continues until the smallest flaw is contained in the moving class width group. The class width is increased by 0.001 inch and the specimens are regrouped using the larger class width and starts at the largest flaw size. The DOEPOD analysis again evaluates the POH and LCL obtained from the flaw data within the larger class width group. This larger

class width group is again decremented (moved) as before until the smallest flaw is contained in the class width group. This process continues for all flaw sizes and class widths until all the flaws are eventually contained within one wide class width group or until the lower confidence bound of any group equals or exceeds 0.90. If a lower confidence bound does (does not) equal or exceeds 0.90 at any class width, then there does (does not) exists a grouping of flaws detected with 0.90 POD with 95% confidence, X_{pod}. If X_{pod} exists, then DOEPOD requires further validation that the POD increases with flaw size (this increase is not assumed a priori) within the range of flaw sizes for which the results are valid. DOEPOD addresses validation at large flaw sizes by using two sequentially applied analyses.

CASE DEFINITIONS

DOEPOD analysis identifies several CASES of data sets depending on the results of the DOEPOD procedures. Selected definitions are abbreviated below,

CASE 1 : The probability of detection meets or exceeds 0.90 with 95% confidence for all flaws at and above the flaw size X_{pod}.

CASE 1*: The probability of detection meets or exceeds 0.90 with 95% confidence at a flaw size X_{pod}. Further evaluation at flaw sizes greater than X_{pod} is required by explaining and resolving Misses above X_{pod}.

CASE 2 : The probability of detection meets or exceeds 0.90 with 95% confidence at a flaw size X_{pod}, however, there are an excessive number of Misses above X_{pod}. Additional evaluation at identified flaw sizes is required.

DOEPOD EXTENDED FOR LARGE FLAWS

Grouping of flaws by number[14, 15] is allowed as long as the four requirements for using binomial statistics are met: (1) The number of trials, N, is to be fixed, (2) Each observation is independent, (3) Each observation represents one of two outcomes (Hit or Miss), and (4) The true probably of Hit is the same for each possible outcome.

In order to meet one of the requirements, the estimated probability of a hit should not be varying substantially within the large flaw size grouping. This is expected to be approximately true when the probability of detection meets or exceed 0.90 with 95% confidence at flaw size, X_{pod}.

Grouping of large flaws by number[14, 15] is executed in DOEPOD analysis when X_{pod} has been identified in a data set. The number of large flaws in any one group is not fixed, and there may be up to 76 adjacent flaws in a group. X_p identifies the minimum flaw size at which all flaw sizes greater or equal to X_p may be grouped by varying the number, N, of flaws in a group to yield a lower confidence bound that meets or exceeds 0.90 with 95% confidence. The procedure for finding X_p is briefly described here.

Starting at the largest flaw size, XL, the 29 largest flaws are put into a group. If the binomial analysis of this group yields a lower 95% bound on POD of 0.90 or greater, then the 90/95 POD is demonstrated conservatively at the largest flaw. If a Miss is observed in the set of 29 largest flaws then the size of the group is expanded to include

the 46 largest flaws. If no additional Misses are observed in the group of 46 samples then 90/95 POD is demonstrated at the largest flaw size. If a second Miss is observed in the set of 46 samples then the group is expanded to the 61 largest flaws. This process is continued until 90/95 POD is demonstrated or a maximum of 76 samples are included in the grouping. Once 90/95 POD is demonstrated at the largest flaw size, this flaw size is labeled X_p,

The next question to answer is whether it is possible to demonstrate 95/95 POD at a smaller size flaw. A new grouping is identified and starts with the second smallest flaw as the candidate. As in the procedure described in the previous paragraph, the second smallest flaw and the next 28 smaller flaws are put into a group of 29. If the binomial analysis of this group yields a lower bound of 0.90 or greater (i.e., there are no misses in the group of 29), then the 90/95 POD is demonstrated at the this smaller flaw size, and the value of X_p is now changed to the second smallest flaw in the data set. Again, if a Miss is observed in this new group of 29, then the group is expanded in size as before. This procedure continues until either $X_p = X_{pod}$ or until the value of X_p can not be made smaller. If $X_p > X_{pod}$, then there is a validation gap between X_{pod} and X_p that can not be resolved with the existing data.

In Figures 1, 2, 7, and 8, the range of flaw sizes where the lower 95% confidence bound on POD meets or exceeds 0.90 is shown as a shaded horizontal bar which extends from the flaw size, X_p, to the largest flaw size, X_L. The presence of X_p indicates that there is 95% confidence that POD meets or exceeds 0.90 for all flaw sizes at and above X_p. If $X_p = X_{pod}$, then the lower bound of the probability of detection meets or exceeds 0.90 with 95% confidence for all flaw sizes from X_{pod} to X_L. It will be shown here that grouping large flaws by number, N, is a necessary but not a sufficient requirement for demonstrating that the lower bound of probability of detection for these groups of large flaws sizes meets or exceeds 0.90 with 95% confidence.

VALIDATING DOEPOD

This work is very specific to validating that the binomial based methodology of DOEPOD yields a 90/95 POD flaw size when it exists and checks for situations where the POD might not be increasing monotonically. Validation is very important in safety-critical applications. The DOEPOD procedures are to provide a validated method where there is no other validated means of establishing POD monotonicity, as it is this 90/95 POD flaw size that is used and reported in formal inspection requirements[8, 17] and investigation documents[18]. The 0.90 value is also used in formal government reports and handbooks[5, 8, 17, 18]. It is especially important to recognize that the POD for specified flaws sizes may or may not meet or exceed 0.90 with 95% confidence depending on the POD methodology used. For example, point estimates and confidence bounds may be obtained by use of parametric-model based methods or nonparametric approaches.[6] Therefore it is important to validate that proposed POD procedures properly identifies when the lower bound of the POD for specified flaws sizes is below, meets, or exceeds 0.90 with 95% confidence. For fracture critical applications there is limited interest in a point estimate value, by itself. However, there is considerable interest in assuring that the POD level exceeds 0.90 with high confidence, and this is the reason that the 90/95 POD flaw size is the sought after level of inspection capability.

The following describes the DOEPOD validation testing performed to demonstrate that DOEPOD identification of the 90/95 POD flaw size, without large flaw misses or false call warnings, qualifies that the inspection system is adequate. That is, if a particular inspection procedure passes the DOEPOD test and an X_{pod} is found, then there is a minimum flaw size, X_{pod}, at which the POD meets or exceeds 0.90 with 95% confidence, and that the POD for range of larger flaws from X_{pod} to X_L also meets or exceeds 0.90 with 95% confidence."

The identification of X_{pod} only indicates that there is a configuration of test samples with adjacent flaws sizes for which the POD, at that point only, meets or exceeds 0.90 POD with 95% confidence. It is emphasized here that even when X_{pod} exists, 90/95 POD at X_{pod} is not demonstrated until it is verified that the POD is at at least 0.90 for flaw sizes greater than X_{pod}.

There are two phases to the validation testing. Phase I is to validate that the DOEPOD analysis identifies a X_{pod} flaw size that is conservative relative to the predicted 90/95 POD flaw size obtained by use of the Logit-ML statistical model. This Phase I validation compares flaw size at X_{pod} with the estimated 90/95 POD flaw size obtained from a parameter based model. Reiterating, DOEPOD analyses do not use a model that implies that the POD is increasing above the X_{pod} flaw size. This is in contrast to the commonly-used parametric models (e.g., the binary regression logit model), for which the POD is monotonically increasing A second validation Phase II is needed to demonstrate that, when X_{pod} exists, X_{pod} is the demonstrated 90/95 POD flaw size. Phase II requires an evaluation of POD at flaw sizes that are greater than X_{pod}. The number and distribution of large flaws needed to make this evaluation is not known. The purpose of Phase II testing is to find the minimum sample requirements for DOEPOD analysis to yield a determination on whether we can be 95% confident that POD exceeds 0.90 for flaws sizes larger than the X_{pod}. A positive determination does not imply that the POD is increasing with flaw size, but rather that the POD is not decreasing at larger flaws.

If the above minimum sample requirements are met and 0.90 POD with 95% confidence is also observed for flaws sizes larger than the X_{pod} flaw size, then X_{pod} is the conservative and demonstrated value for the 90/95 POD flaw size as determined by the binomial applications in the DOEPOD methodology.

The current implementation of DOEPOD v.1.0 requires 25 flaws with flaw sizes greater than X_{pod} and equally distributed in sizes up to and including the largest flaw size of the test set. The origin of this requirement will be identified in the following Phase II validation. In order to demonstrate that the effect of varying the number of large flaws, and to verify that this requirement is adequate, the DOEPOD analysis for large flaws needs to be turned off or inhibited. If the DOEPOD analysis for large flaws is executed, then the results will always indicate an exception for sample sets with less than 25 large flaws.

Phase I: Validate 90/95 POD at X_{pod} is a Conservative Value

A DOEPOD analysis was performed on each of the 437 POD data sets in NTIAC NDE Capabilities Data Book. Results of these DOEPOD analyses identifies whether X_{pod} exists in each data set. 153 of the 437 data sets are identified to be CASE 1 or CASE 1* and yield 90/95 POD flaws sizes, by DOEPOD binomial distribution method, and the Logit-ML procedures[7], respectively. These are the 153 data sets that have 90/95 POD points that may be compared between these two POD methods. Further, for 145 of the 153 data sets, DOEPOD analysis yields an observed X_{pod} flaw size that is conservative (i.e., larger flaw size) when compared to the 90/95 POD flaw size provided by the Logit-ML method. That is, DOEPOD yields a conservative value of the X_{pod} flaw size when a 90/95 POD flaw size is also estimated using the Logit-ML procedures, and that this is true for 95% of the data sets compared. For the other eight of 153 data sets, DOEPOD analysis yields an observed X_{pod} flaw size that is at least 15% smaller than the 90/95 POD flaw size provided by the Logit-ML method. . The 15% difference is chosen to define and quantify a significant difference between the observed X_{pod} flaw size and the 90/95 POD flaw size obtained from the estimated two parameter statistical model.

A careful examination of the eight data sets that exhibit X_{pod} flaw sizes that are significantly smaller than the 90/95 value provided by Logit-ML identifies both data integrity issues and/or inadequacy of the two parameter statistical model. One of the eight data sets has erroneous analysis in the NTIAC NDE Capabilities Data Book. When the Logit-ML analysis is corrected, DOEPOD analysis yields a X_{pod} flaw size that is larger than the estimated 90/95 POD flaw size obtained from the Logit-ML method. One of the eight data sets contains mixed sample thicknesses for an analysis by crack depth to thickness ratio. Comparisons of this data set with other data sets analyzed by either crack length or crack depth is not appropriate for this validation. There are two data sets, of the eight, for which the estimated 90/95 POD flaw sizes obtained from the Logit-ML method are outside the range of the actual flaw sizes in the data set. Use of the estimated 90/95 POD flaw size for these two data sets without supporting test data near the estimated 90/95 POD flaw size, implies extrapolation is not good engineering practice. This highlights the potential risk of improper use of the POD curve fitting procedures.

As a result, there are only four data sets out of 437, where DOEPOD analysis yields CASE 1 or CASE 1* with an observed a X_{pod} flaw size that is more than 15% smaller than the 90/95 POD flaw size provided by the Logit-ML method.

Further evaluation of the four data sets exhibiting an X_{pod} flaw size that is less than the 90/95 POD flaw size provided by the Logit-ML method reveals that the logit model does not fit the estimated probability of Hit proportions (POH) from the observed data very well. This lack of fit is quantitatively identified by large standard errors, shown in Table 1, between the Logit-ML predicted POD and the observed probability of Hit proportions (POH). A quantitative comparison between good and poor curve fits is discussed below.

Table 1

Data Set[7]	Root Mean Square Deviation Between POD and Logit-ML
D7002L	0.1814
D7001L	0.1878
CA003(3)L	0.1319
G2001L	0.2097

Table 1. Root Mean Square Deviation Between POD and Logit-ML estimates exhibiting X_{pod} flaw sizes that are less conservative with respect to the estimated 90/95 POD flaw sizes obtain from the two parameter statistical model.

Analysis results for the D7002L data set shown in Figure 1 highlights the rather poor fit of the logit model (upper dashed curve), as measured by the Root mean square deviation between POD (observed proportions, POH shown as open circles) and Logit-ML, 0.1814. Here DOEPOD analysis identifies and observed X_{pod} flaw size (upper most solid triangle) of 0.066". In comparison to the Logit-ML 90/95 POD flaw size of 0.165". The POH proportions are from flaws all having sizes within 0.020" of each other so these grouped flaws are similar in size.

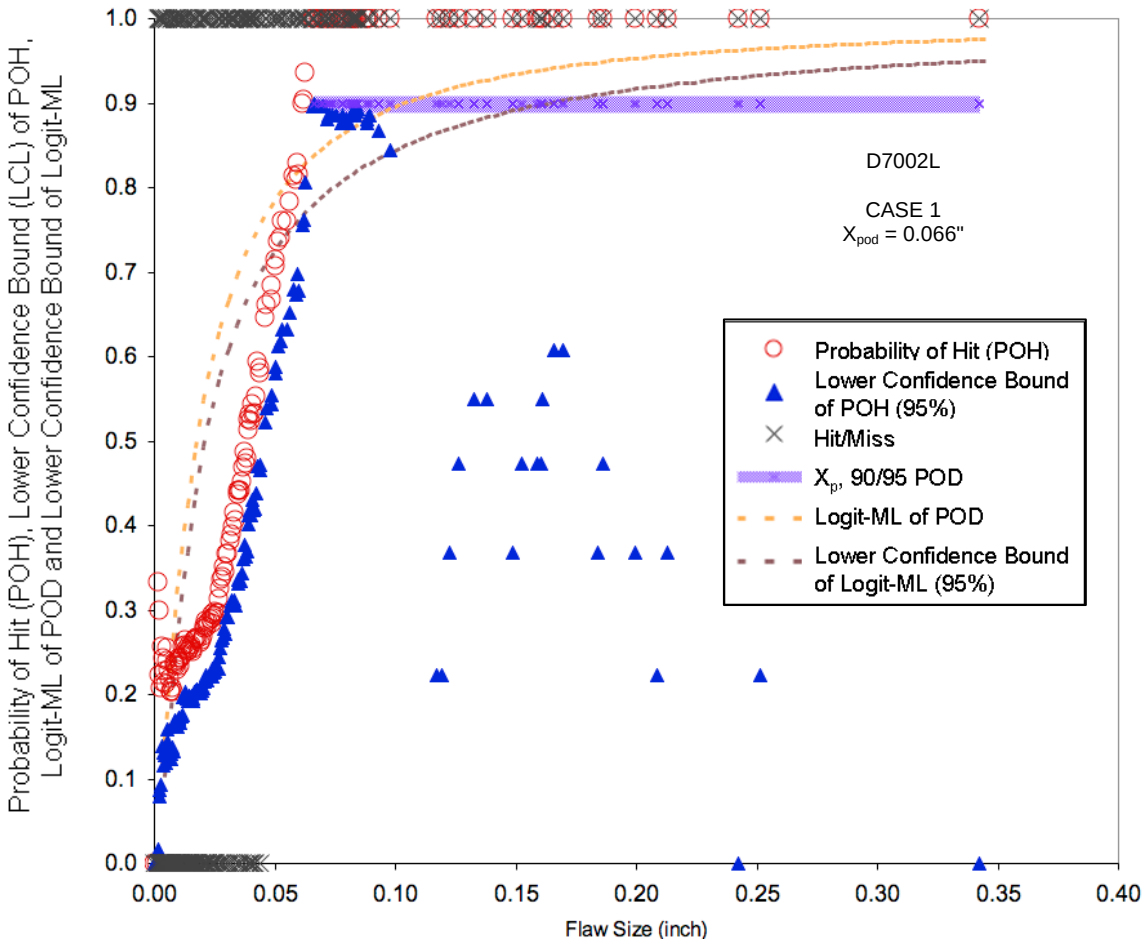

Figure 1. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for data set D7002L.

Analysis results of data set A8002L are shown in Figure 2 for comparison, where the root mean square deviation between POD and Logit-ML is small, 0.08, and the Logit-ML estimates track the observed proportions (POH) well. Here DOEPOD analysis identifies the X_{pod} flaw size (upper most solid triangle) at 0.0147 inches in comparison to 0.0103 inches, the 90/95 POD flaw size provided by Logit-ML. In this example, the 90/95 POD values from the Logit-ML and DOEPOD analysis are similar and within 0.004 inches of each other (the example in Figure 1 had a 0.1 inch difference between values).

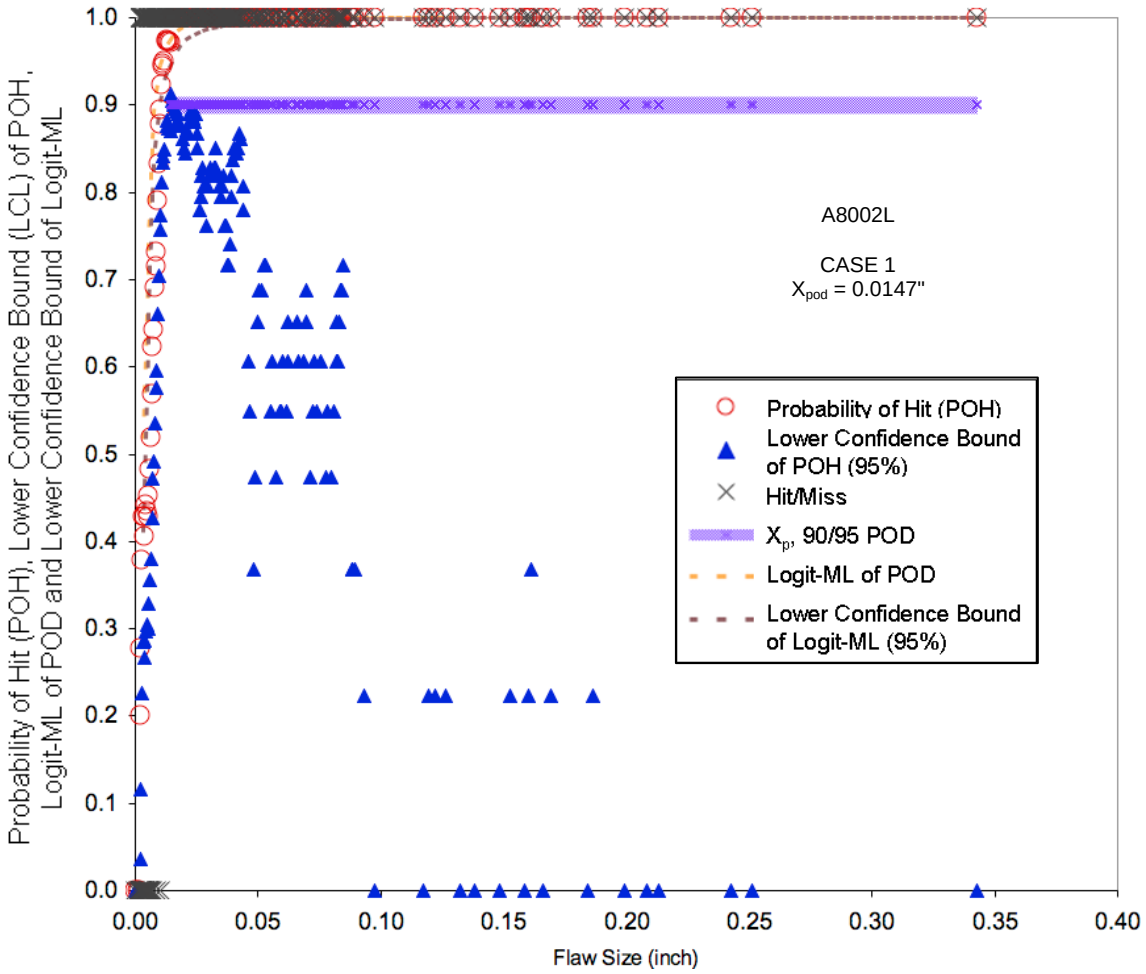

Figure 2. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for data set A8002L.

Summarizing the above Phase I results. When the binomial-based analysis used in DOEPOD identifies a CASE 1 or CASE 1* data set exhibiting a X_{pod} flaw size, this flaw size is usually larger than (and thus conservative relative to) the Logit-ML POD 90/95 POD flaw size. Exceptions arise when the logit model does not fit the observed data well. This is shown to be true even when large flaw number and size distribution requirements are not specified. This lack of fit exception occurred in four of the 153 (CASE 1 and CASE 1*) data sets in the NTIAC NDE Capabilities Data Book, 1997 where X_{pod} and 90/95 POD are obtained via DOEPOD analysis and Logit-ML, respectively. The Logit-ML POD method is inadequate for at least these four data sets. It is re-emphasized here that even when DOEPOD analysis identifies X_{pod} to exist at a flaw size, 90/95 POD at X_{pod} and for all larger flaws sizes is not demonstrated until it is verified that the POD is at an acceptable level for flaw sizes greater than X_{pod}.

Phase II: Determining Sample Requirements for Large Flaws

The question to be answered in this section is: "If only the tested flaws are those that lead to the 90/95 POD being met or exceeded at X_{pod} only, what additional flaws are needed to assure that 90/95 POD is also met or exceeded for larger flaws?" This may be

also stated as: "What additional large flaws are required to demonstrate that the POD is monotonic above X_{pod}." In an effort to answer these questions some background and concerns on past proposed methods for making this determination will be discussed. A Monte Carlo evaluation of selected existing data sets will provide an answer to the question above. The Phase II section relies entirely on real data and a series of discrete steps are taken to assure that selected data sets are adequate to serve as the domain for the Monte Carlo evaluation.

Background and Concerns on Establishing Monotonicity

It has been shown that for the existing data sets discussed, the X_{pod} flaw size provided by DOEPOD from CASE 1 or CASE 1* data sets is conservative relative to the to the Logit-ML POD 90/95 POD flaw size, except when logit model does not fit the observed data well. One important aspect of relying on the flaw size as determined by a 29/29 confidence bound (or equivalent) test is that it still remains unknown whether the POD is increasing with increasing flaw sizes above the identified flaw size where a lower bound meets or exceeds 0.90 with 95% confidence is at that point only. The POD also needs to be evaluated at larger flaw sizes. The determination of CASE 1 and CASE 1* for the data sets above was made without having the number of large flaws and the large flaw size distribution requirements being specified. Interestingly, even with this lack of specification the DOEPOD analysis yielded conservative values of X_{pod} with respect to Logit-ML 90/95 POD flaw size. This assurance of conservativeness is false since the comparison is being made with another methodology that may be inadequate. It will be shown here that, for real world applications, the number of large flaws and the large flaw size distribution requirements must be specified.

When the 90/95 POD is not met or exceeded at large flaw sizes, DOEPOD analysis identifies this scenario as CASE $2^{(1, 2, 3, 4)}$, where further data and evaluation is needed for flaws larger than the X_{pod} flaw size. In prior work[1, 2, 3, 4] it was suggested that validation at larger flaw size may be performed by at least three different methods.

The first method is to repeat the 29/29 confidence bound (or equivalent) testing at two additional flaws sizes: (1) at the largest flaw size in the data set, and (2) at a flaw size midway between X_{pod} and the largest flaw size in the data set. This approach added two additional flaw sizes for which the lower 95% confidence bounds for POD may be demonstrated to meet or exceed 0.90. Unfortunately, the required number of inspected specimens to make two additional statements is not available in existing data sets where there are limited samples at the mid-point and largest flaw sizes. This does not mean that testing at a mid-point and the largest flaw size is inadequate, but rather that these demonstrations do not allow for direct comparison with existing data sets. Given the lack of real data with the structure needed to make the additional two statements, an alternate approach needs to be explored for determining the large flaw requirements from existing data. The second method is to include an additional 27 flaws at equally distributed sizes between X_{pod} and largest flaw size of the test set, and subsequently grouping of flaws by number. A challenge presents itself in identifying what number and acceptable distribution of flaw sizes are to be evaluated above the X_{pod} flaw size. A Monte Carlo testing approach that utilizes existing data sets to estimate the large flaw requirements is pursued in this work. The third method is the development of procedures for using good engineering judgment supported by data obtained from similar systems.

There is also a caution noted here when identifying flaw sizes for all POD studies. Selection of flaw sizes may be dependant on physics of the inspection system. For example, if a differential eddy current probe system is being evaluated and if the flaw sizes are greater than the eddy current footprint, then there is a possibility that the POD will decrease when the flaw size is greater than the eddy current footprint. Flaw sizes that address these issues need to be included in the POD test.

Identification of Adequate Existing Data Sets for Monte Carlo Evaluation

The fist step in identifying adequate data sets is to select data sets that have an identified X_{pod} flaw size and that have excessive Misses above that flaw size. The presence of these two features reveals a data set where one might erroneously claim that the X_{pod} flaw size is the 90/95 flaw size. The DOEPOD procedures identify all data sets with these features as CASE 2 data sets. CASE 2 data sets all have excessive Misses for flaws larger than $X_{pod.}$ This CASE 2 designation is made when the binomial analysis of numbers of grouped large flaws[14, 15] (flaws sizes greater than the X_{pod} flaw size) results in demonstrating that 90/95 POD is not met or exceeded for all flaws sizes greater than the X_{pod} flaw size. This binomial analysis of numbers of grouped large flaws is a quantitative evaluation of Misses that are outliers.

DOEPOD identified 46 CASE 2 data sets out of 437 POD data sets in the NTIAC NDE Capabilities Data Book where further data and evaluation is needed to validate that the 90/95 POD is met or exceeded for all flaws sizes greater than the X_{pod} flaw size.

The second step is to select CASE 2 data sets where X_{pod} that is less than the logit-ML 90/95 POD flaw size. These data sets represent the most risk. If only the flaws in the small grouping at X_{pod} were evaluated, so that no large flaw evaluation is performed, then X_{pod} maybe erroneously claimed to be the 90/95 POD flaw size. In contrast, if the CASE 2 data sets exhibit a X_{pod} that is equal to or greater than the logit-ML 90/95 POD flaw size, then one might argue that X_{pod} is simply a conservative value of the 90/95 POD flaw size. Therefore, in an effort to reduce risk, it is prudent to focus on potential high risk scenarios where X_{pod} that is less than the logit-ML 90/95 POD flaw size.

12 of the 46 CASE 2 data sets yield an observed X_{pod} flaw size that is at least 15% smaller (i.e., less conservative) with respect to the logit-ML 90/95 POD flaw size. Note that X_{pod} represents only one point and not the 90/95 POD flaw size. The 12 data sets where these non-conservative CASE 2 scenarios occur are A3001BL, A6003H, B1003AL, C6003AL, C8001(3)D, CE011(6)D, CE011(6)L, D1002BD, D8001(3)L, D8003(3)D, D8003(3)L, and DC002(3)D. These 12 data sets represent possible data samplings for which a single 29/29 confidence bound (or equivalent) test may result in a X_{pod} flaw size that is not conservative with respect to Logit-ML 90/95 POD flaw size. That is, if the initial specimen or data set is a selected subset of the entire specimen or data set, then a single point estimate may lead to an apparent 90/95 POD flaw size that is non-conservative, with respect to the Logit-ML 90/95 POD flaw size. Or more directly what if only these selected specimens where generated and tested, then the results of the test on the larger flaws remains unknown, and unknown risk is introduced.

This risk is highlighted in the next two charts. The DOEPOD and Logit-ML analyses of the original D8001(3)L full data set is shown in Figure 3.

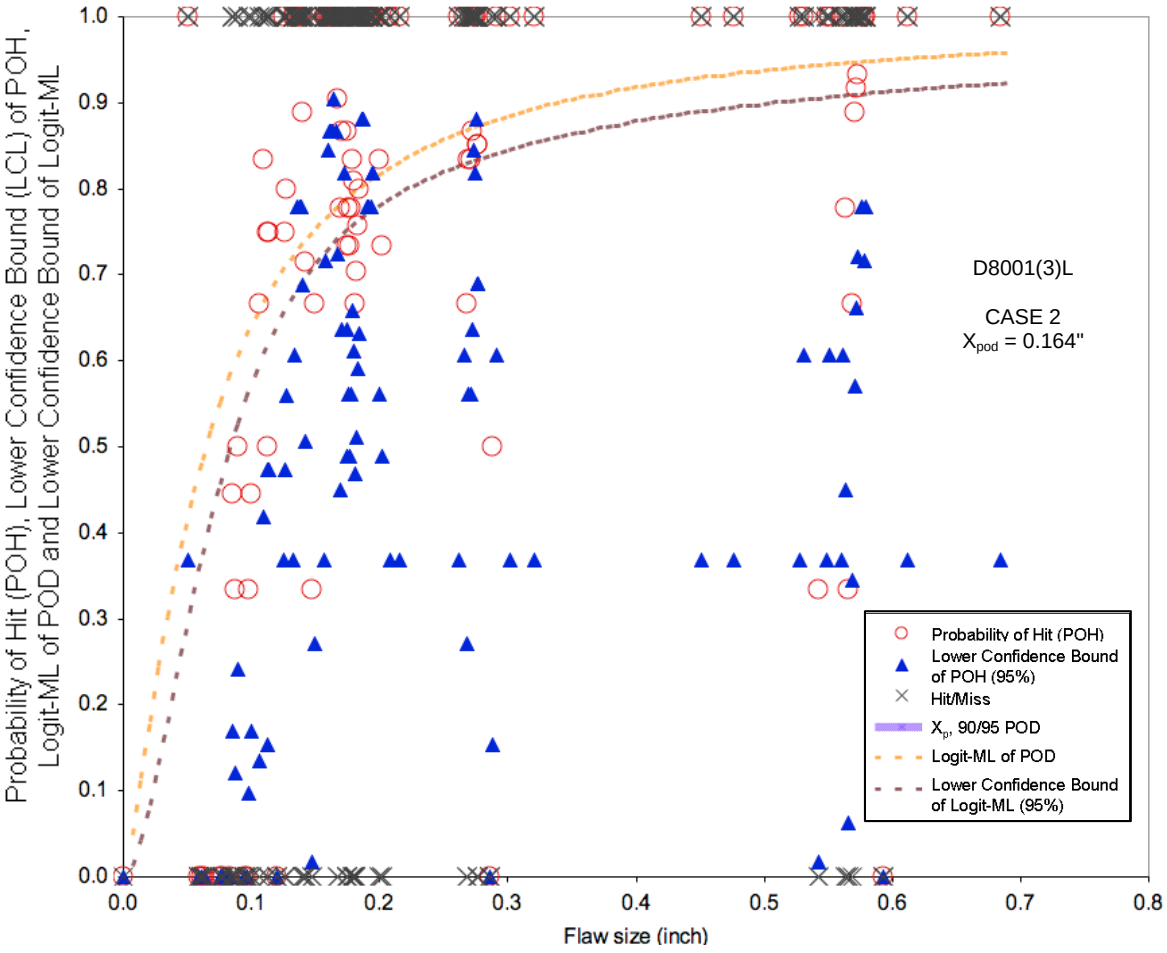

Figure 3. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for data set D8001(3)L.

The X_{pod} flaw size (upper most solid triangle) for this full data set is 0.164 inches. In contrast, by selecting a small sample consisting of a subset of the original D8001(3)L data, one may obtain an identical X_{pod} flaw size, as shown in Figure 4. Here this subset contains only flaws with sizes less than or equal to 0.164 inches.

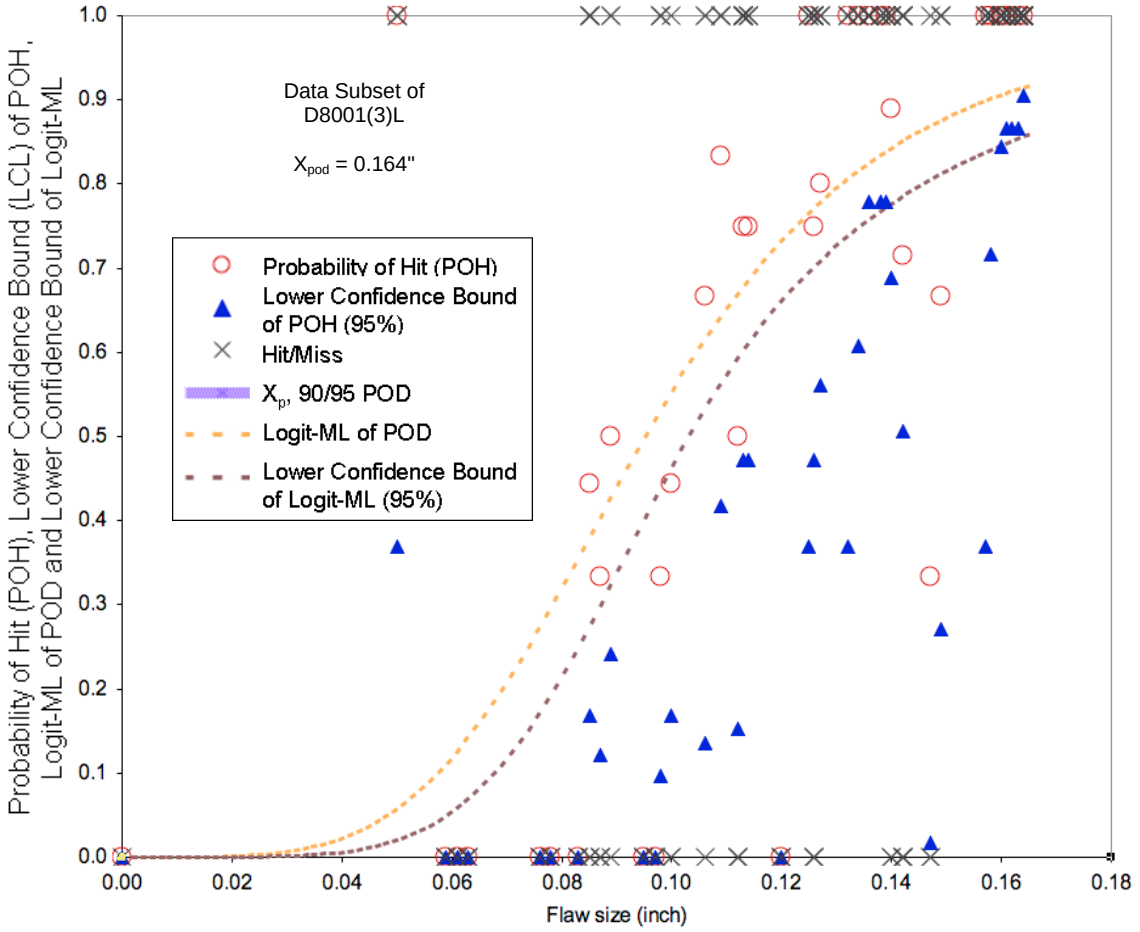

Figure 4. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for a subset of data in D8001(3)L.

The DOEPOD analysis yields an X_{pod} flaw size (upper most solid triangle) at 0.164 inches for both the data sets shown in Figures 3 and 4. However, 90/95 POD is not demonstrated for either the full data set or subset of data because 90/95 POD is not met or exceeded for all flaws sizes greater than the X_{pod} flaw size. Remembering here that, by the minimum sample requirements, the largest flaw size must be equal to or greater than $3xX_{pod}$ in order to establish that 90/95 POD is met or exceeded for all flaws sizes greater than the X_{pod} flaw size. The Logit-ML curve fitting procedure shows the predicted POD (upper dashed curve) increasing for all flaws sizes and for both data sets, however, the presence of 10 Missed (out of 62 opportunities) large flaws above 0.510 inches in the original data set, makes this Logit-ML predicted POD questionable and suggests that an alternative model is required. One possibility would be a model in which the POD asymptotes to something less than 1 for large flaws.

This information now provides us with guidance on how to proceed in validating that the POD is actually increasing with flaw sizes greater than the X_{pod} flaw size. First, the lower bound on POD must meet or exceed 0.90 with 95% confidence must be reached at some flaw size, X_{pod}. Second, a range of flaw sizes above X_{pod} needs to be included in the data set. Third, the parameter based predictive POD models should not be relied upon for demonstrating that the POD is increasing with flaw size above the 90/95 POD flaw size. That is, the adequacy of the predictive model is not assured.

The third step in identifying data sets that are adequate is to select data sets that have a sufficient number and range of large flaw sizes available above the X_{pod} flaw size so that uniquely different samples can be selected. There should be no Misses at the largest flaw size. By DOEPOD design, CASE 1 can never occur when there is a Miss at the largest flaw size, so that data sets containing these features are excluded.

Using all of the above constraints, there are two (2) original CASE 2 data files from which to generate random data files for the Monte Carlo test. The files are labeled as A6003H and D1002BD.

Monte Carlo Procedure

The testing domain is the data from identified files (A6003H and D1002BD) in the NTIAC NDE Capabilities Data Book. Subsample data files are randomly generated from the two domain files. The DOEPOD analysis is performed on the individual subsample data files. The individual DOEPOD analysis results are aggregated into a final result.

Generating Subsample Data Sets

In order to perform this Monte Carlo evaluation, a series of randomly generated subsample data files are required where the number of flaws having sizes greater than the X_{pod} flaw size is allowed to increase from 2 to 35. The number range is arbitrary where the actual number required is, at this point, unknown. Data sets can be generated by a "Delete-by-M" jackknife[10] subsampling method, where M denotes the total number of large flaws excluded from the original data set. There is no replacement of samples, so that once a sample has been randomly selected, that sample can not be selected again for the same subsample data set. By dynamically changing M, the sensitivity of the DOEPOD procedures to properly determine that 90/95 POD is met or exceeded for all flaws sizes greater than the X_{pod} flaw size is explored as a function of the number of large flaws. By construction, the "Delete-by-M" procedure is only applied to add to the data set flaws with sizes greater than X_{pod} flaw size.

The first Monte Carlo subsample data set contains only two large flaws is generated by randomly selecting one sample having a flaw size greater than the X_{pod} flaw size. The largest flaw in the original data set is also included. The largest flaw in the original data set is included to define the upper limit of the flaw size range. All flaws are drawn from the original real data set. . This completed Monte Carlo subsample data set now contains all the original flaw sizes up to the X_{pod} flaw size and one additional randomly selected flaw larger than the X_{pod} flaw size and the largest flaw in the original data set. As with many statistical tests, increasing the number of Monte Carlo subsample data sets, generally increases the confidence of the results. 76 complete random individual Monte Carlo subsample data sets are generated by repeating the above process 76 times. The 76 subsample data sets comprise one complete collection of randomly generated input data files containing only two large flaws in each subsample date set. The process is repeated for 2, 3, 4, … , 34 randomly selected large flaws sizes to yield a total 2584 randomly generated input data files. The range of the number of selected large flaws is chosen to reflect the range of minimum number of large flaws (two) to a

maximum number of large flaws (thirty-four) available in the data sets. A total of 5168 random subsample data files are generated for the A6003H and D1002BD data sets.

DOEPOD Analysis Results

DOEPOD analysis results from the 5168 subsample data sets are used to specify the large flaw requirements needed to assure that proper determination is made on whether or not 90/95 POD is met or exceeded for all larger flaw sizes. There are two possible outcomes from the DOEPOD analysis of the randomly generated subsample files. The DOEPOD analysis yields CASE identifications that are either a failure or a success. Since these original data sets are CASE 2 data sets, DOEPOD analysis should not identify the subsample data sets as CASE 1 data sets. Therefore, for validating the conservative nature of the DOEPOD analysis, a failure is defined as CASE 1 (i.e., 90/95 POD is met or exceeded for all flaws sizes greater than the X_{pod} flaw size.). A success is defined as any other CASE (i.e., 90/95 POD is not met for all flaws sizes greater than the X_{pod} flaw size.). Here the presence of CASE 1 represents a failure in the DOEPOD analysis for any of the subsample data sets, and for either of the original CASE 2 data sets A6003H and D1002BD, and this failure represents added risk. By varying M, the minimum number of large flaws required to assure a conservative determination of the true CASE may be obtained.

The original D1002BD CASE 2 and A6003H CASE 2 data sets are shown in Figures 5 and 6, respectively.

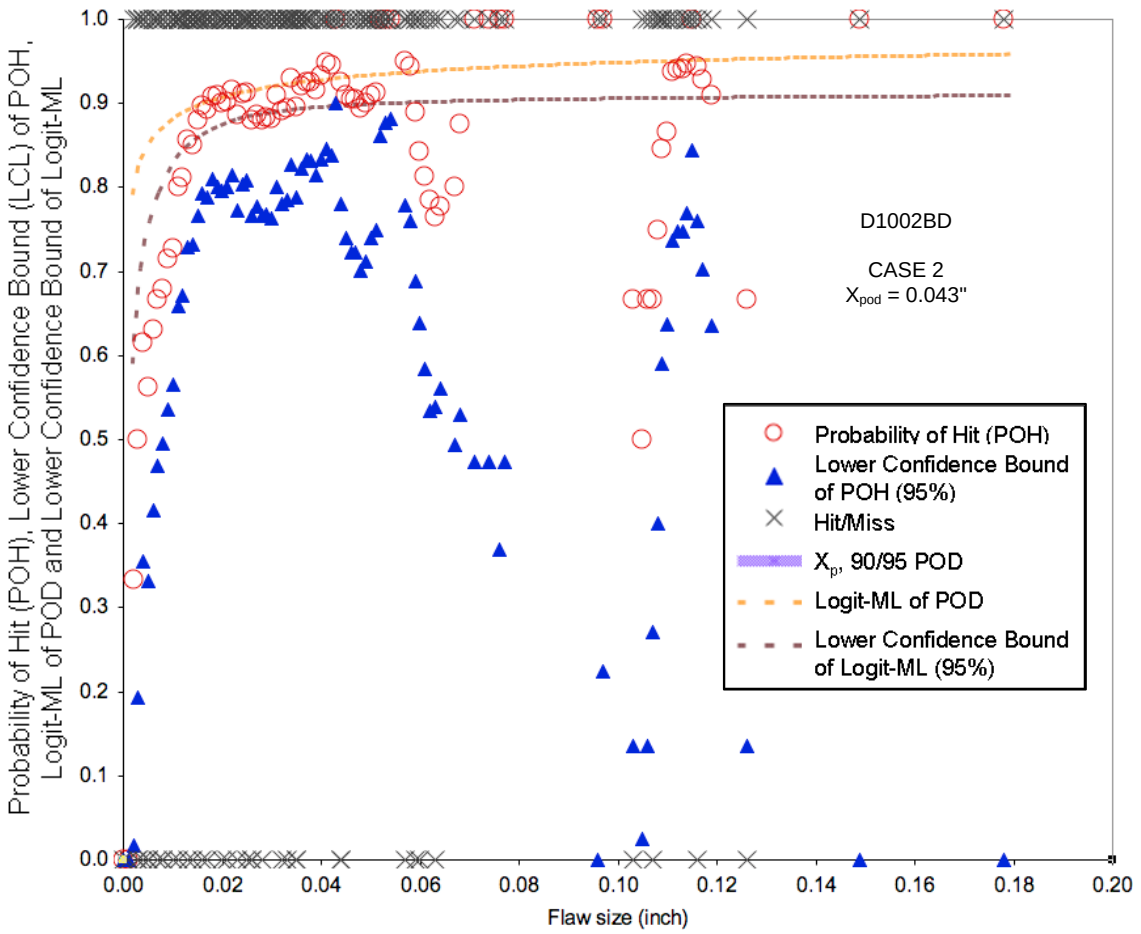

Figure 5. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for data set D1002BD.

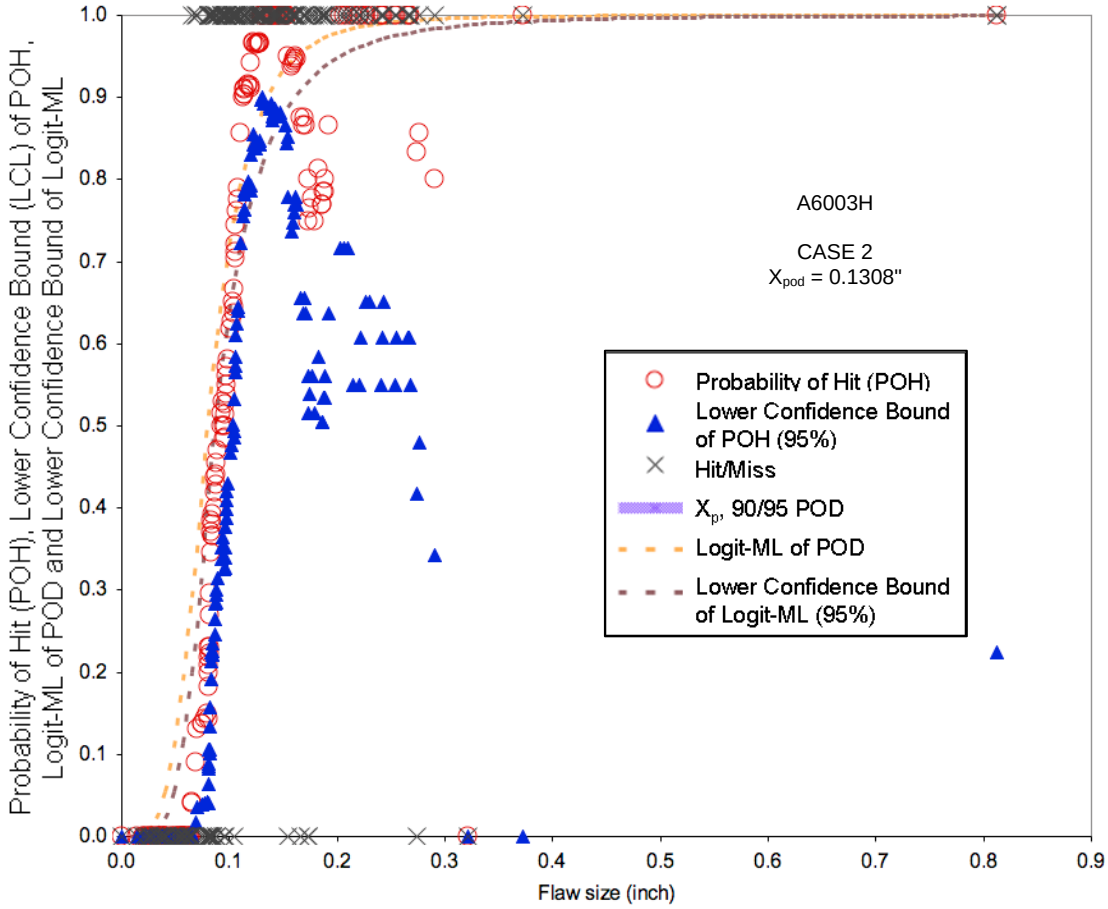

Figure 6. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size for data set A6003H.

The examination of two typical Monte Carlo generated data sets highlights the risk of using data sets with a limited numbers of large flaws.

A typical random data set generated from the original D1002B data is shown in Figure 7 for trial number 68 when 20 larger flaws are randomly selected for this trial. DOEPOD analysis yields a CASE 1* and is a success, i.e., not CASE 1, because there are conditions on CASE 1* that limit the validation at large flaw sizes. The conditions are that Misses must be explained and resolved before validation at large flaw sizes is accepted. This is the DOEPOD analysis indication that 90/95 POD is not met or exceeded for flaws greater than X_{pod} and more evaluation is required. The additional evaluation here must address the requirement that every Miss observed at flaw sizes greater than X_{pod} must be explained and resolved. That is, the inspection documents, the inspection procedure, and the physical integrity of the flaw are to be verified. If the inspection procedure is inadequate, e.g., the sample was not cleaned properly, or if there was a data recording error, then the Miss maybe explained and retesting is warranted. These types of exceptions are typical when they do occur.

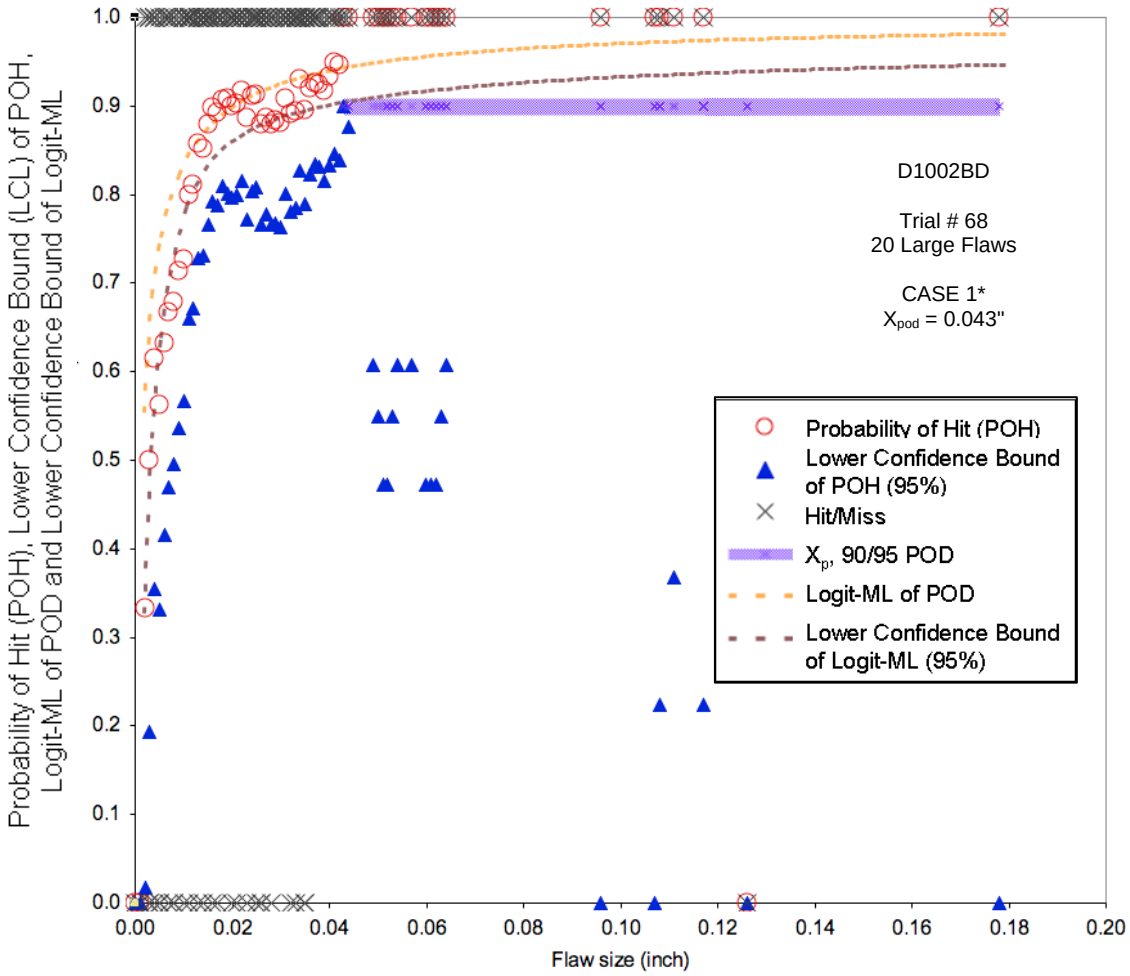

Figure 7. Trial #68 with 20 random large flaws from data set D1002BD. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size.

In contrast, another typical random data set generated from the original D1002BD data is shown in Figure 8 for trial number 65 when 20 larger flaws are randomly selected for this trial. DOEPOD yields a CASE 1 and is a failure, since there are no conditions on CASE 1 that limit the validation at large flaw sizes. Note the absence of inspection Misses in this random data set above the X_{pod} flaw size, 0.043 inches. This trial represents added risk where the random data selected from the original CASE 2 data set yields a CASE 1. That is, the DOEPOD analysis of this random Monte Carlo test data subsample does not identify any difficulty in detecting large flaws, even when 20 large flaws are included in the analysis. If this were the only inspection data taken, the result would erroneously imply that 90/95 POD is met or exceeded at and above X_{pod}, and this represents increased risk.

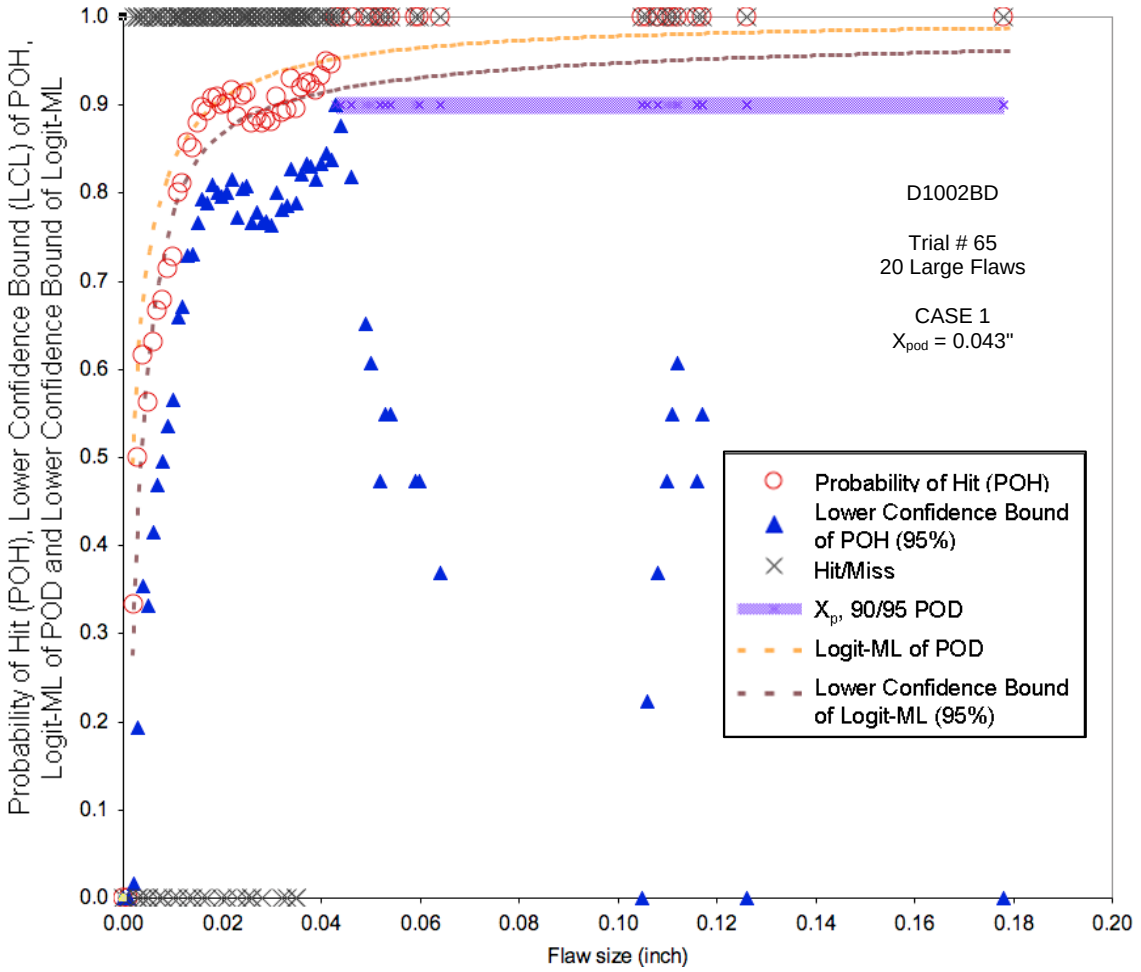

Figure 8. Trial #65 with 20 random large flaws from data set D1002BD. Probability of Hit (POH), POH Lower Confidence Bound (LCL), Logit-ML of POD and Lower Confidence Bound of Logit-ML versus flaw size.

Aggregating the Individual DOEPOD Analysis Results

The individual DOEPOD analysis results for the 5187 subsamples are aggregated into a final result by evaluating the estimated probability of success (POS) that DOEPOD procedures determine POD to be non-monotonic above the X_{pod} flaw size when it is non-monotonic. POS is estimated as a function of the number of randomly selected large flaws.

POS is estimated by applying binomial statistics to the results of each data set having the same number of randomly selected and unique large flaws. As stated earlier, the use of binomial statistics requires that four elements be true if a statistical variable is described by a binomial distribution: (1) The number of trials, N, is to be fixed. N = 76, is the number of runs of DOEPOD. (2) Each observation, i.e., DOEPOD analysis result on a randomly generated data set, is independent, (3) Each observation (DOEPOD analysis result) represents one of two outcomes (success or failure). Any result other than CASE 1 is a success. A CASE 1 result is a failure, and (4) The true probability of success (POS) that DOEPOD identifies a success (i.e., any case other than CASE 1) is the same for each possible outcome for fixed M. The truth of the last element is dependent on the range and distribution of large flaws available in the data sets. It will be shown

later that the coefficient of variation of the large flaw distribution may be used to establish the presence of weighting in a data set.

In this Monte Carlo evaluation there are 76 data sets with the same number of randomly selected flaws for each of the original two data sets (A6003H, D1002BD), or 76 trials with either a failure (CASE 1) or a success (any case other than CASE 1). The ratio

$$POS = \frac{Number\ of\ Sucesses}{Number\ of\ Trials}$$

is a proportion and is an estimate of the probability that the DOEPOD analysis successfully identifies the probability of detection to be non-monotonic for large flaws. The lower bound (LCL) on POS at 95% confidence is also determined[1]. Using the same nomenclature, a 90/95 POS indicates that the there is 0.90 probability of successfully identifies that the POD is non-monotonic with 95% confidence when it is non-monotonic.

A summary of the DOEPOD analysis for both sets of 2584 random data files is shown in Figure 9. The POS exhibits a different structure between the two data sets and this is expected since the distribution of large flaw sizes between the two data sets are different. It is noted here that the proportion given by the ratio of (Number of Inspection Misses)/ (Number of Available Large Flaws) in both A6003H and D1002BD data sets are similar at 0.10 and 0.11, respectively.

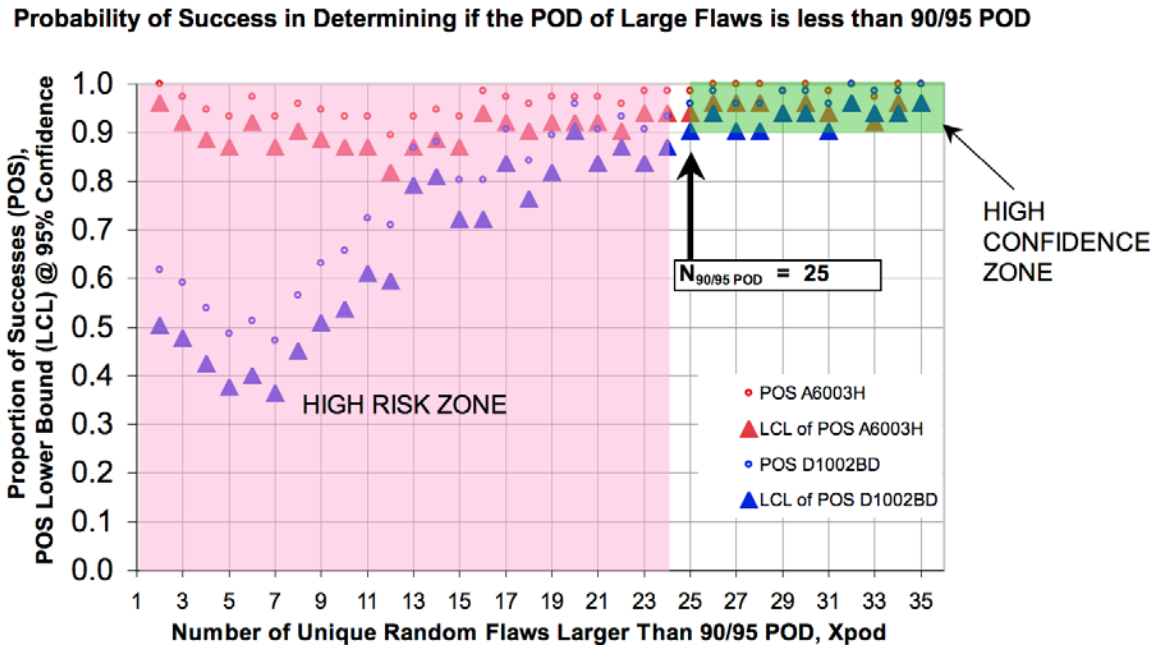

Figure 9. Summary of the DOEPOD analysis for both sets of 2584 random data files. The minimum number of larger flaws, $N_{90/95\ POD} = 25$, required to demonstrate that there is a 0.90 Probability of Success (POS)with 95% confidence that DOEPOD analysis establishes that 90/95 POD is not met or exceeded for large flaws. HIGH CONFIDENCE ZONE: When 0.90 POS with 95% confidence is met or exceeded, LCL is equal to or greater than 90%. The number of flaws, N = 25, (with sizes larger than the X_{pod} flaw size) required to demonstrate that 90/95 POD is met or exceeded for flaw sizes larger than the X_{pod}. This number of large flaws is required when new NDI or enhanced NDI technologies are being evaluated. HIGH RISK ZONE: When 0.90 POS with 95% confidence is not demonstrated, LCL is less than 0.90. The number of

large flaws is insufficient to demonstrate that 0.90 POD with 95% confidence is met or exceeded for flaw sizes larger than the X_{pod}. This number of larger flaws may be accepted, with justification, when conventional or derivative NDI technologies are being evaluated.

The POS is a function of number of large flaws and indicates the number of large flaws required for the DOEPOD analysis to establish whether or not the POD for large flaws meets or exceeds 0.90 with 95% confidence. The number of large flaws required to demonstrate that 90/95 POS is met or exceeded varies for these two CASE 2 examples. Remembering that when 90/95 POS is met or exceeded, the DOEPOD analysis will also properly identify that 90/95 POD is not met or exceeded for large flaws. From Figure 9 the minimum number of large flaws required to establish that 0.90 POD with 95% confidence is met or exceeded for these large flaw sizes and simultaneously for both A6003H and D1002BD data sets may be identified as 25. This occurs when POS is 0.96 with a lower bound of 90%. The standard error on POS is 0.043. From Figure 9, the lower bound on POD for large flaws may be checked by adding 25 ($N_{90/95\ POD} = 25$) or more random and unique flaws with flaw sizes exceeding X_{pod}. Adding 25 or more random and unique flaws with flaw sizes exceeding X_{pod} represents a successful large flaw evaluation test in the HIGH CONFIDENCE ZONE shown in the Figure 9. This test should be considered as mandatory for all evaluations of new or enhanced NDI technologies.

The largest number of large flaws where 0.90 POS with 95% confidence is not met in either A6003H and D1002BD data sets establishes a number of large flaws that, when 90/95 POD is not met or exceeded for large flaws, then the DOEPOD analysis will not adequately identify that 0.90 POD with 95% confidence is not met or exceeded for large flaws. This represents added risk (HIGH RISK ZONE) shown in the Figure 9. Therefore, adding less than 25 random flaws with flaw sizes exceeding the X_{pod} should only be considered when justification is provided and when evaluating conventional or derivative NDI technologies.

A different trend is observed in the POS between the A6003H and D1002BD data sets. The origin of the difference is identified by examination of the distribution of large flaw sizes. A6003H has large flaws grouped together and not uniformly spaced above the X_{pod} flaw size, 0.1308 inches. In contrast, the D1002BD data set exhibits a fairly uniform distribution of large flaws distributed above X_{pod} flaw size, 0.043 inches.

In order to provide the most general stringent test for validation at large flaw sizes, it is appropriate to identify data sets similar to D1002BD as the preferred large flaw size distribution. That is, flaws above X_{pod} need to be uniformly distributed in sizes between the X_{pod} and largest flaw size. The definition of "uniformly" is subjective, however, the coefficient of variation, CV, may be used to test for degree of the uniformity distribution. CV is the ratio of the standard deviation of flaw sizes greater than X_{pod} to the mean of the flaw sizes greater than X_{pod},

$$Coefficient\ of\ Variation,\ \ CV \equiv \frac{Standard\ Deviation\ of\ Large\ Flaws\ Sizes}{Mean\ of\ Large\ Flaws\ Sizes}$$

The DOEPOD methodology provides guidance on the acceptable values of CV. An acceptable range is defined here to have large flaws with sizes approximately equally spaced from X_{pod}, to the largest flaws size, X_L. Data sets with a CV less than 0.33 are not

sufficiently uniformly distributed and exhibit narrow groupings of flaws. When uniformly spaced flaws are considered, a CV of 0.34 is identified as the acceptable for D1002BD, while the actual CV for this data set is 0.39. Large flaws with a CV greater than 0.56 are not sufficiently uniform and exhibit skewed groupings of flaws. This CV is observed for the data set A6003H, while the acceptable CV when considering uniformly spaced large flaws for this data set is 0.40. An examination of the entire set of data files in the NTIAC Capabilities Data Book yields the acceptable CV to be in the range 0.337 – 0.506.

The requirements for 25 unique and uniformly distributed large flaws yielding a CV in the range of 0.33 – 0.51 for these large flaws is a requirement to reach CASE 1 in DOEPOD v.1.0. This requirement assures the probability (POS) that DOEPOD analysis will determine that the POD function is non-monotonic when it is non-monotonic to be 0.96 (96 in a hundred), and therefore, DOEPOD v.1.0 analysis only identifies CASE 1 when in the high confidence zone shown in Figure 9. This requirement is establish by using existing real data. It will be shown later that by using simulations, the Monte Carlo uncertainty on the proportion of successes in determining non-monotonicity is substantially lower when compared with the above 0.043 uncertainty on the POS that is present when the POS is determined by using the limited experimental data from the two real data sets.

Summarizing the above Phase II results: For these two real data sets, a minimum of 25 uniquely different flaw sizes larger than X_{pod} that uniformly span the range from X_{pod} to the largest flaw size, are required for validating that 90/95 POD is met or exceeded in the range from X_{pod} to the largest flaw size. Since a minimum number of flaws needed to identify X_{pod} is 29, and with the requirement for 25 flaws larger than X_{pod}, then the minimum number of flaws required to demonstrate that 90/95 POD is met or exceed at and above X_{pod} is 54 flaws when using the binomial point estimate methodology of DOEPOD. The above requirements were established by generating 5168 subsample data sets from only two real data sets. It is noted here that this sub-sampling, and subsequent analysis, only explores the variability in the estimators, X_p and number of large flaws needed, that are used in DOEPOD, for these particular data sets. It is assumed, for now, that these requirements are adequate for all POD data sets. The "Delete-by-M" jackknife[10] sub-sampling method was used to generate these samples. Monte Carlo simulations to follow will further demonstrate that this assumption is correct.

DOEPOD v.1.0 evaluates the number and distribution of flaws with sizes greater than X_{pod} for validating that 90/95 POD is met or exceeded in the range from X_{pod} to the largest flaw size.

ESTABLISHING THE PROOF PROPERTY BY MONTE CARLO SIMULATIONS

The prior validations were done using existing real inspection data sets. The advantage to using real data sets is that shape of the POD function is not assumed. Use of existing data sets allowed for the determination of the minimum number and distribution of large flaws sizes needed to identify 90/95 POD flaw size. However, since the number of real data sets that meet the test requirements is limited, it is prudent to explore the application of these large flaw requirements to selected POD functions that

simulate possible variations from assumed monotonicity. The goal of these simulations is to establish the proof property of the DOEPOD procedures for determining whether the POD function is either non-monotonic or monotonic.

Test Procedure Using Functions that Simulate POD

The procedure used for simulations is straight forward. (1) Select POD functions that are to simulate different types of POD curves. These may be monotonic, non-monotonic, as well as oscillating. (2) Create 2000 data sets by randomly drawing samples according to the Hit-Miss proportions at selected flaw sizes. The flaw sizes chosen are determined by the requirements of the DOEPOD procedures. The flaws are randomly dithered in sizes about the flaws sizes required by the DOEPOD procedures in order to represent the distribution of real flaw sizes. (3) Analyze each simulated data set following the DOEPOD procedures. The analysis results may indicate that more samples are needed. If more samples are needed, then these samples are drawn until the sample requirements are met or a Miss is obtained, at which time the updated data set in re-analyzed using the DOEPOD procedures. This process continues until a determination may be made as whether the POD function is monotonic (CASE 1) or non-monotonic (all other CASEs). (4) Aggregate the results to determine the capability of the DOEPOD procedures.

In order to constrain the number of analysis cycles need to make a determination of monotonicity, only 29 executions of the DOEPOD procedures are allowed. The flaws size resolution is constrained when evaluating whether $X_{pod} = X_p$ is true or not. Here X_{pod} is considered to be equal to X_p when $X_p - X_{pod} < 0.002$ inches. In order to avoid including flaws at excessively large sizes, the large flaw range can not be extended more than three times. For example, if the largest flaw size, XL, is 0.7 inches in the initial data set, then the DOEPOD procedures will not continue to request flaw sizes greater than 5.6 inches.

There are two different modes that will be used for establishing the capabilities of the DOEPOD procedures. The first mode is an iterative sequential looping mode that starts with only 16 flaws. After the iterative process is complete and a determination on whether the POD is monotonic or non-monotonic is made, the total number of samples will have increased. The average number of samples needed (ASN), and the standard deviation of the ASN, SD, for the 2000 initial sample sets is recorded for evaluating the efficiency of the DOEPOD procedures. The second mode is a non-looping mode that uses only fifty-four (54) flaws. The fifty-four (54) sample requirements have been determined by the prior DOEPOD results from real data sets. The fifty-four (54) samples are to contain (twenty-nine) 29 flaws at the target 90/95 POD flaw size and twenty-five (25) larger flaws uniformly spaced in size between the target 90/95 POD flaw size and three times the target 90/95 POD flaw size. This mode is used to address the question: If a 29/29 result is observed at the target flaw size, what is the capability of the DOEPOD procedures to make a determination on whether the POD is monotonic or non-monotonic?

POD Simulation Functions

Representative POD functions are selected to examine the characteristics and capability of the DOEPOD procedures in determining whether the POD function is non-monotonic or monotonic. One way to determine these characteristics is to use POD functions that have both subtle and dramatic non-monotonicity.

Six different POD functions provide both subtle and dramatic variations are developed and shown in Figure 10. The POD functions labeled B,C, D, and E all are non-monotonic and asymptote to 0.90, 0.85, 0.75, and 0.50 POD, respectively. The POD function labeled A is non-monotonic and exhibits a 13.5% oscillatory dip from a maximum value such that a minimum occurs at 0.85 POD and the POD function asymptotes to 1.00 for the largest large flaws. The POD function labeled G is monotonic and is the basis from which all other POD functions have been derived.

POD functions A, B, C, D, E are all evaluated using the iterative sequential looping mode starting with survey set 16 samples have flaws sizes of 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7 inches, while the POD functions G and C are evaluated using the non-looping mode and starting with 54 samples that have 29 flaws at the target flaw size of 0.29 inches and 25 flaws uniformly spaced in size extending from 0.29 inches to 0.9 inches. During sample selections the flaw sizes are randomly dithered about the flaw sizes listed above.

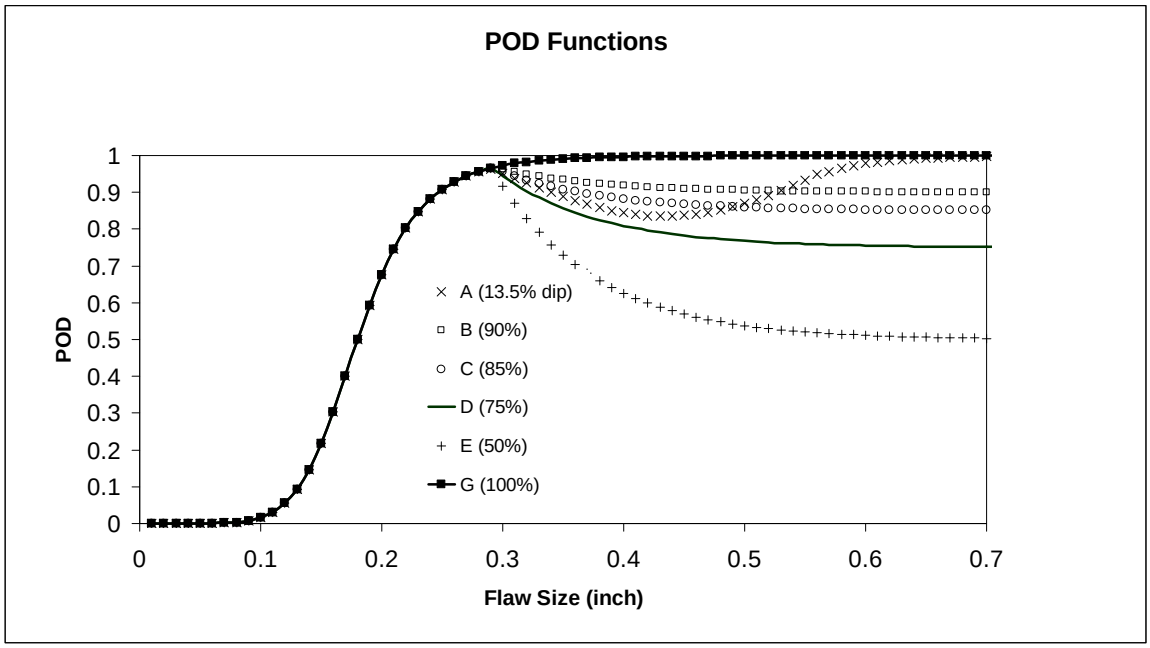

Figure 10. Selected POD functions used in Monte Carlo tests.

The first five rows of Table 2 give results using DOEPOD in the sequential mode in which additional flaws are added to the sample set in order to resolve ambiguity. With

true POD curve C in Figure 10, the probability that the DOEPOD analysis will indicate that the POD function is monotonic when it is not is 0.004 (4 in a thousand). For true POD curve B, the probability is 0.053 (53 in 1000).

The last two rows of Table 2 give results using DOEPOD in a single run mode with 54 samples (which would be used when it is impossible to add additional flaws to the sample set). With true POD curve C in Figure 10, the probability that will indicate that the POD function is monotonic when it is not is 0.028 (28 in 1000). On the other hand, with true POD curve G in Figure 10, which is monotone increasing, the probability that will indicate that the POD function is monotonic is 0.949 (949 in 1000).

The Monte Carlo standard errors on the proportion of non-monotonic determinations, p, are small and do not exceed 0.0098.

TABLE 2

ID	POD FUNCTIONAL FORM	ASYMPTOTE OF POD FUNCTION	PROPORTION OF MONOTOINC, (1-p)	PROPORTION OF NON-MONOTONIC, p	MONTE-CARLO ERROR ±1.96[p(1-p)/2000]1/2	AVERAGE SAMPLE NUMBER, ASN	STANDARD DEVIATION OF ASN, SD
A	Oscillating (13.5% dip)	0.850	0.040	0.960	0.0086	267	79.9
B	Non- Monotonic	0.900	0.053	0.947	0.0098	91.1	56.3
C	Non- Monotonic	0.850	0.004	0.996	0.0028	59.7	27.1
D	Non- Monotonic	0.750	0.000	1.000	0.0000	37.7	12.6
E	Non- Monotonic	0.500	0.000	1.000	0.0000	25	5.1
C	Non- Monotonic	0.850	0.028	0.972	0.0072	54	0
G	Monotonic	1.000	0.949	0.051	0.0096	54	0

Table 2. Summary of Monte Carlo tests that utilize the simulation functions A, B., C, D, E, and G.

The average sample number (ASN) and its standard deviation, SD, for each of the simulations are also shown in Table 2. The ASN for the single run non-looping mode are fixed at 54 as indicated in the last two rows in table 2.

When using DOEPOD in the sequential iterative mode the ASN and SD increase as the non-monotonic POD function approaches a monotonic function. The non-monotonic POD function that has an asymptotes to 0.50 (E) requires considerably smaller ASN than the non-monotonic POD function that asymptotes to 0.90 (B). When the non-monotonic and monotonic functions become similar, it takes more data in order to make a determination of monotonicity. This trend is further evidenced by the large ASN (267) needed to make a determination on monotonicity for the POD function with a slight dip (A).

The frequency of occurrence versus the number of samples required to make a determination on montonicity, when DOEPOD analysis is used in the sequential iterative

mode is shown in Figure 11. The distributions of the number of samples are uniform for POD function, B, C, D, and E. However, the number of samples has a bimodal distribution for the POD function A. The origin of the bimodal shape of the number of samples is likely to be due to the fact that, although the POD function (A) is non-monotonic below 0.524 inches, this POD function is monotonic above 0.524 inches.

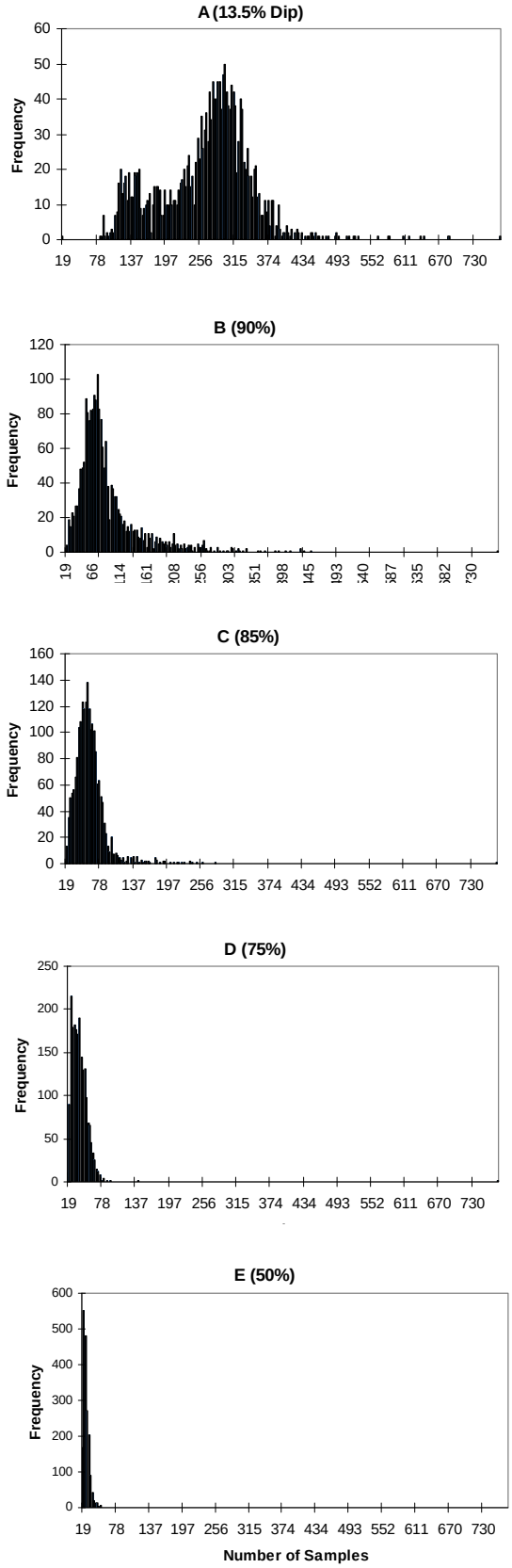

Figure 11. The frequency of occurrence versus the number of samples required to make a determination on monotonicity when DOEPOD analysis is used in the sequential iterative mode.

The DOEPOD analysis reveals this monotonicity by identifying 90/95 POD flaw sizes shown in Figure 12. 90/95 POD flaws sizes smaller than 0.524 inches represents a 4% error where the DOEPOD procedures determined this POD function to be monotonic below 0.524 inches when it is not. In contrast, the DOEPOD procedures make a determination that the POD function is monotonic at and above 0.524 inches where it is monotonic.

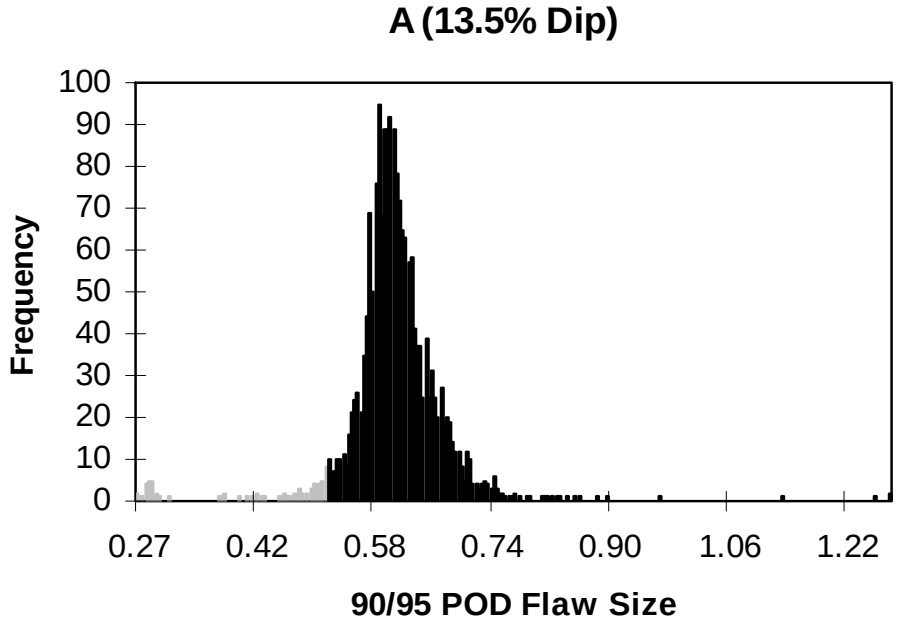

Figure 12. The 90/95 POD flaw sizes for simulations based on the POD function A. The light shaded region represents where the DOEPOD procedures determined this POD function to be monotonic below 0.524 inches when it is not. The darker shaded region represents where the DOEPOD procedures make a determination that the POD function is monotonic at and above 0.524 inches where it is monotonic.

SUMMARY

The DOEPOD Phase I and II validation testing demonstrates that the DOEPOD methodology yields a demonstrated and conservative estimated value of the 90/95 POD flaw size with respect to 90/95 POD flaw size obtained by maximum likelihood estimation of the two parameter Logit statistical model[7] and confidence bound procedures[9] used in the NTIAC NDE Capabilities Data Book[7] when adequate statistical models are used.

A minimum of 25 randomly selected uniquely different flaw sizes larger than the X_{pod} flaw size, that uniformly span the range from X_{pod} to the largest flaw size, that are required for validating that 0.90 POD with 95% confidence is met or exceeded in the range from X_{pod} to the largest flaw size. Since a minimum number of flaws needed to

identify X_{pod} is 29, and with the requirement for 25 flaws larger than X_{pod}, then the minimum number of flaws required to demonstrate that 0.90 POD with 95% confidence is met or exceed at and above X_{pod} is 54 flaws when using the binomial point estimate methodology of DOEPOD.

Monte Carlo simulations validate that the procedures used in DOEPOD identify a non-monotonic POD function at least 95 $\pm$ 0.98 % of the time when it exists.

If a 90/95 POD flaw size is established using 29 similar flaws, then with the inclusion of 25 uniformly spaced larger flaws the DOEPOD procedures will yield a determination that the POD is monotonic 95 $\pm$ 0.96% of the time when it is monotonic.

If a 90/95 POD flaw size is established using 29 similar flaws, then with the inclusion of 25 uniformly spaced larger flaws the DOEPOD procedures will yield a determination that the POD is non-monotonic 97 $\pm$ 0.72% of the time when it is non-monotonic.

REFERENCES

(1) E. R. Generazio, Directed Design of Experiments for Validating Probability of Detection Capability of NDE Systems (DOEPOD), 34th Annual Review of Progress in Quantitative Nondestructive Evaluation, Golden, Colorado, July 22-27, 2007.

(2) Review of Progress in Quantitative Nondestructive Evaluation, Volume 975, pgs 1693-1700, American Institute of Physics, Melville, New York, 2008.

(3) E. R. Generazio, Design of Experiments for Validating Probability of Detection Capability of NDT Systems and for Qualification of Inspectors, Materials Evaluation, Vol. 67, No. 6., pg 730-738, and Errata, Materials Evaluation, Vol.67, No. 6, September 2009, pg 1041.

(4) United States Patent Application 20100122117, Directed Design of Experiments for Validating Probability of Detection Capability of a Testing System.

(5) Mil-HDBK-5H, Metallic Materials and Elements for Aerospace Vehicle Structures, December 1, 1998

(6) G. J. Hahn and W. Q. Meeker, Statistical Intervals: A Guide for Practitioners, Wiley, 1991

(7) NDE Capabilities Data Book, 3rd ed., Nov. 1997, NTIAC DB-97-02

(8) Mil-HDBK-1823 Nondestructive Evaluation System Reliability, 30 April 1999

(9) NDE Detectability of Fatigue-Type Cracks in High Strength Alloys, Martin Marietta, MCR-88-1044, September 1988

(10) Delete-m Jackknife for Unequal m, Frank M. T. A. Busing, Erik Meijer, and Rien Van Der Leeden, Statistics and Computing (1999) Vol. 9, pgs. 3-8

(11) An Empirical Distribution Function for Sampling with Incomplete Information, Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman, Annals of Mathematical Statistics, Vol. 26, No. 4 (1955), pgs. 641-647.

(12) W. W. Hines and D. C. Montgomery, Probability and Statistics in Engineering and Management Science, 2nd Edition, 1980, John Wiley & Sons

(13) Ewout W. Steyerberg , Clinical Prediction Models, A practical Approach to Development, validation, and Updating, Springer Science and Business Media, 2009

(14) B. G. W. Yee, et al., Assessment of NDE Reliability Data, NASA-CR-134991, Oct. 1976

(15) Ward D. Rummel, Recommended Practice for Demonstration of Nondestructive (NDE) Reliability on Aircraft Production Parts, Materials Evaluation, vol 40, August 1982

(16) Standard NDE Guidelines and Requirements for Fracture Control Programs, MSFC-STD-1249, National Aeronautics and Space Administration, September 1985.

(17) NASA-STD-5009, "Nondestructive Evaluation Requirements for Fracture Critical Metallic Components", April 7, 2008.

(18) Aircraft Accident Report, Uncontained Engine Failure, Delta Airlines Flight 1288, McDonnell Douglas MD-88, N927DA, Pensacola, Florida, July 6, 1996, January 13, 1998, National Transportation Safety Board, Washington, D.C. 20594, NTSB/AAR-98/01

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01- 09 - 2011	Technical Publication	

4. TITLE AND SUBTITLE

Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Generazio, Edward R.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

724297.40.44.07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

L-20056

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)

NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

NASA/TP-2011-217176

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 38
Availability: NASA CASI (443) 757-5802

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The capability of an inspection system is established by applications of various methodologies to determine the probability of detection (POD). One accepted metric of an adequate inspection system is that for a minimum flaw size and all greater flaw sizes, there is 0.90 probability of detection with 95% confidence (90/95 POD). Directed design of experiments for probability of detection (DOEPOD) has been developed to provide an efficient and accurate methodology that yields estimates of POD and confidence bounds for both Hit-Miss or signal amplitude testing, where signal amplitudes are reduced to Hit-Miss by using a signal threshold Directed DOEPOD uses a nonparametric approach for the analysis or inspection data that does require any assumptions about the particular functional form of a POD function. The DOEPOD procedure identifies, for a given sample set whether or not the minimum requirement of 0.90 probability of detection with 95% confidence is demonstrated for a minimum flaw size and for all greater flaw sizes (90/95 POD). The DOEPOD procedures are sequentially executed in order to minimize the number of samples needed to demonstrate that there is a 90/95 POD lower confidence bound at a given flaw size and that the POD is monotonic for flaw sizes exceeding that 90/95 POD flaw size. The conservativeness of the DOEPOD methodology results is discussed. Validated guidelines for binomial estimation of POD for fracture critical inspection are established.

15. SUBJECT TERMS

Probability of detection; Failure critical; Fracture critical; Nondestructive testing; Nondestructive evaluation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	38	(443) 757-5802