NATIONAAL LUCHT- EN RUIMTEVAARTLABORATORIUM

NATIONAL AEROSPACE LABORATORY NLR

THE NETHERLANDS

NLR TR 84011 U

RELIABILITY OF NONDESTRUCTIVE INSPECTION

BY

J.H. HEIDA

NLR

DOCUMENT CONTROL SHEET

	ORIGINATOR'S REF.		SECURITY CLASS.
	NLR MP 84011 U		Unclassified

ORIGINATOR

National Aerospace Laboratory NLR
Amsterdam, The Netherlands

TITLE

Reliability of nondestructive inspection

PRESENTED AT

AUTHORS	DATE	pp	ref
J.H. Heida	840207	30	17

DESCRIPTORS

Reliability
Inspection
Nondestructive tests
Probability
Confidence limits
Intervals
Ultrasonic flow detection
Maximum likelihood estimates
Thresholds (perception)
Sensitivity
Quadrinomial distribution
Accuracy
Defects
Regression analysis
Statistical tests

ABSTRACT

This report describes the concept of reliability of nondestructive inspection (NDI) and reviews the different aspects involved.

First, the probability of detection (POD) of flaws in flawed specimens is discussed. To provide a measure of confidence in an estimated POD, lower bound values of this probability at a given confidence level are calculated with statistical methods. The inspection sample size in practice limits the applicability of statics. Therefore the available data points are grouped in intervals of flaw size; four interval methods are discussed in this report. Also a different approach for the representation of NDI reliability, by making use of linear regression analysis, is described.

Next, the more general quadrinomial distribution of the possible outcomes of NDI, with successful and unsuccessful inspections of both flawed and unflawed specimens, is discussed. In analog with the POD for the flawed specimens, a probability of recognition (POR) is defined for the unflawed specimens. Both POD and POR must be included in a characterization of NDI reliability. Possible measures for this characterization are discussed. It is concluded that it is doubtful whether a single parameter will give an appropriate measure of NDI reliability; instead it is recommended to establish minimum values for both the POD and POR at a specified confidence level.

Finally, some remarks are given about the application of reliability demonstration programs for the determination of flaw detection capabilities of various NDI-methods.

NLR TR 84011 U

RELIABILITY OF NONDESTRUCTIVE INSPECTION

by

J.H. Heida

SUMMARY

This report describes the concept of reliability of nondestructive inspection (NDI) and reviews the different aspects involved.

First, the probability of detection (POD) of flaws in flawed specimens is discussed. To provide a measure of confidence in an estimated POD, lower bound values of this probability at a given confidence level are calculated with statistical methods. The inspection sample size in practice limits the applicability of statistics. Therefore the available data points are grouped in intervals of flaw size; four interval methods are discussed in this report.
Also a different approach for the representation of NDI reliability, by making use of linear regression analysis, is described.

Next, the more general quadrinomial distribution of the possible outcomes of NDI, with successful and unsuccessful inspections of both flawed and unflawed specimens, is discussed. In analogy with the POD for the flawed specimens, a probability of recognition (POR) is defined for the unflawed specimens. Both POD and POR must be included in a characterization of NDI reliability. Possible measures for this characterization are discussed. It is concluded that it is doubtful whether a single parameter will give an appropriate measure of NDI reliability; instead it is recommended to establish minimum values for both the POD and POR at a specified confidence level.

Finally, some remarks are given about the application of reliability demonstration programs for the determination of flaw detection capabilities of various NDI-methods.

Division: Structures and Materials
Prepared: JHH/
Approved: GB/

Completed : 7-II-1984
Ordernumber: 542.003
Typ. : MM

CONTENTS

(30 pages in total)

1 INTRODUCTION

The concept of reliability of nondestructive inspection (NDI) is an often discussed subject and it has been defined in numerous ways. This report will give a review of the current knowledge of the different aspects involved and of the relevant literature.

Generally, the reliability of a NDI technique is associated with the probability of detection (POD) of a flaw of a particular size as function of that size. The true POD for a particular flaw can, however, only be obtained by means of an infinite number of inspections; in practice a limited number of inspections will only yield an estimated POD. To provide a measure of confidence in the estimated POD, it is usual to calculate so-called confidence limits with statistical methods, resulting in lower-bound values of the POD. Reliability of NDI can therefore be better described in terms of POD and the degree of confidence in that POD. Section 2 of this report will discuss different aspects involved, such as the calculation of the confidence limits, the significance of the probability/confidence mix, and the number of inspections required for a certain reliability of inspection. Also a different approach for the representation of NDI reliability, by making use of linear regression analysis, will be described.

In practice, however, also specimens without flaws are inspected, with the probability of obtaining a spurious indication of a non-existing flaw. The results of NDI can hence be described by means of a quadrinomial distribution, with successful and unsuccessful inspections of both flawed and unflawed specimens. In analogy with the POD for the flawed specimens a probability of recognition (POR) will be defined for the unflawed specimens. Both POD and POR are essential inspection characteristics with their relative importance depending on considerations of safety and economics. Possible measures for the characterization of NDI reliability will be discussed in section 3 of this report, by making use of the fraction of detected flaws and the fraction of false calls in the quadrinomial distribution.

To demonstrate the flaw detection capabilities of various NDI methods, so-called "reliability demonstration programs" are conducted. In section 4 of this report some remarks will be given about such programs.

2 PROBABILITY OF DETECTION AND ITS CONFIDENCE

2.1 Probability of detection curve

The reliability of a nondestructive inspection (NDI) technique is generally associated with the probability of detection (POD) of a flaw of a particular size as function of that size. This can be illustrated with the aid of a POD curve, see figure 1. The curve is always dependent on the different inspection conditions, such as the applied NDI method and the specimen configuration. It will be clear that the POD is a function of flaw size. Below a certain threshold value of the flaw size the POD will be negligible. This threshold value is a measure of the sensitivity of the applied NDI method. For large flaw sizes the POD will approach 100 percent. In the intermediate flaw range the POD curve will show an increasing, nonlinear, behaviour.

In first instance, most attention was focused in the NDI world on the sensitivity of a NDI method, i.e. the smallest detectable flaw size. When the concept of reliability was introduced, however, it became clear that of far more importance is the largest flaw size that is missed sometimes by an inspection technique.

Another concept frequently used in relation to NDI is the accuracy of an inspection technique. Accuracy can be defined as the degree of correspondence between the flaw size emerging from an inspection and its actual size. Although an essential part of NDI, the concept of accuracy and the application of defect sizing techniques, will not be further discussed in this report.

Further it is emphasized that in section 2 of this report only the detection of flaws in flawed specimens is considered. In section 3, the more general quadrinomial distribution, including the treatment of false calls, will be discussed.

2.2 Confidence limits

Each point of the POD curve in figure 1 is theoretically based on an infinite number of inspections. In practice, however, for a particular flaw size we will only have a limited number of inspections, yielding an estimated POD, the so-called point estimate $\bar{p}$:

$$\bar{p} = n/N \quad ; \; n = \text{number of detections}, \; N = \text{number of inspections.} \qquad (1)$$

It is to be expected that the point estimate will differ from the true probability p:

$$p = \lim_{N \to \infty} (n/N). \qquad (2)$$

The difference between $\bar{p}$ and p will become smaller for increasing number of inspections. The question now arises what degree of certainty or what confidence we can have in the estimated POD if we only have a limited number of inspections. Because of the statistical variation in $\bar{p}$, confidence limits can be calculated with statistical methods, resulting in a lower bound value of $\bar{p}$, also called the lower confidence limit p_{ℓ}.

Starting point for the calculation of confidence limits is the assumption that the number of detections for a flaw with particular size (a) has a binomial distribution. This is correct because there are only two possible outcomes for the inspection of that flaw: a yes-or-no detection (hit or miss). Sometimes the Normal, Poisson or Chi-square distributions are used as approximations to the binomial distribution; their applicability will not be discussed in this report but can be found in any textbook on statistical methods. In reference 1 methods and formulae are given for the computation of lower bound values for the point estimate $\bar{p}$ at a given confidence level.

Assume that p is the true (but yet unknown) probability of detection, as defined in (2). The true probability of a miss q then equals (1-p) because of the binomial character of the inspection result (p+q=1). The probability that there will be n detections in a total number of N inspections of a particular flaw size, assuming that the true POD is p, is given by the binomial or Bernoulli probability function:

$$P(X=n) = \binom{N}{n}.p^{n}.q^{N-n}; \text{ with } \binom{N}{n} = \frac{N!}{n!(N-n)!} \qquad (3)$$

where the random variable X denotes the number of detections in N inspection; n = 0,1,....,N. The sum of all possible values of P in eq. 3 is equal to unity and can be written as:

$$\sum_{x=0}^{N} \binom{N}{x}.p^{x}.q^{N-x} = (p+q)^{N} = 1. \qquad (4)$$

The probability of detecting n <u>or</u> more flaws can be found by summing eq. 3 over all values of X for which $X \geq n$. Thus,

$$P(X \geq n) = \sum_{x=n}^{N} \binom{N}{x} . p^x . q^{N-x} . \qquad (5)$$

To provide a measure of confidence in the estimated POD, the point estimate $\bar{p} = n/N$, it is now usual to calculate a lower bound value of $\bar{p}$, i.e. a lower confidence limit p_ℓ, such that there is a high degree of confidence that the true probability of detection p is greater than or equal to p_ℓ. For this purpose a confidence level G, or the fraction $\alpha = 1 - G$, is introduced. The requirement of a high degree of confidence G such that $p \geq p_\ell$ can also be described as the requirement that the chance of detecting n or more flaws, considering a binomial distribution with probability of detection p_ℓ, is less than or equal to $\alpha = 1 - G$ (see Fig. 2):

$$\sum_{x=n}^{N} \binom{N}{x} . p_\ell^x . (1-p_\ell)^{N-x} \leq 1 - G. \qquad (6)$$

Equation 6 can be interpreted as follows. With a single (k) binomial experiment (N, n_k) the corresponding point estimate $\bar{p}_k$ and, with equation 6, the lower confidence limit $p_{\ell k}$ can be calculated. If this binomial experiment is repeated an infinite number of times, the true probability of detection p will then be the average of the individual point estimates $\bar{p}_k$. For all experiments the corresponding values of $p_{\ell k}$ are calculated with equation 6, of these values 100 G % will have a value less than or equal to p. We can now say with 100 G % confidence that the $p_{\ell k}$ of any specific binomial experiment will be less than or equal to p. Said otherwise: with 100 G % confidence we can say that the true probability of detection p is greater than or equal to the lower confidence limit $p_{\ell k}$, as estimated with equation 6 from a single binomial experiment.

The choice of G is arbitrary and is determined by the desired level of confidence for the probability of detection p not to be situated outside the interval range of p_ℓ to 1. It is not difficult to see that if larger values of G are chosen (closer to 1), that the corresponding values of p_ℓ will become smaller. For example if no misses are allowed out of 30 inspections, at the confidence levels of 90, 95, 99, 99.5 percent the resulting lower confidence limits p_ℓ are 92.6, 90.5, 85.8 and 83.8, respectively.

The lower confidence limit p_ℓ can alternatively be calculated, given n, N and G, by making use of the F distribution. Reference 2 (page 697 ff) gives the following equation for p_ℓ:

$$p_\ell = \frac{n}{n + (N-n+1).F_G(f_1,f_2)} \qquad (7)$$

where $F_G(f_1,f_2)$ is the appropriate percentile of the F distribution, with the specified level of confidence G and the degrees of freedom f_1 and f_2.

$$f_1 = 2(N-n+1)$$
$$f_2 = 2n. \qquad (8)$$

In table 1 an example is given of the 95th percentile values for the F distribution ($F_{0.95}$, confidence level 95 %) with two degrees of freedom f_1 and f_2.

2.3 Probability/confidence mix

The flaw detection capability of a nondestructive inspection method is generally described in terms of a specified combination of the probability of detection and the degree of confidence in that probability, also called the probability/confidence mix. Often, a 90/95 mix is quoted for reliable inspection of flaws. For example, in the USAF Airplane Damage Tolerance Requirements (MIL-A-83444, Ref. 3) and in the USAF Damage Tolerant Design Handbook (Ref. 4, in fact written to provide specific background data and justification for the requirements of MIL-A-83444) a probability/confidence mix of 90/95 is required for the reliable inspection in slow crack growth structures of all flaws larger than a specified size. This means that for those flaws it has to be shown that there is a 95 % confidence that the true POD is 90 % or more. Said otherwise: there is only a 5 % probability that the true POD is less than 90 %. The 90/95 mix was specified in reference 4 for flaws in slow crack growth structures, i.e. "structures which are designed such that initial damage will grow at a stable, slow rate under service environment and not achieve a size large enough to cause rapid unstable propagation". For the category of fail-safe structures, i.e. "structures which are designed such that propagating damage is safely contained by failing a major load path or by other damage arrestment features", a probability/confidence mix of 90/50 was specified in reference 4.

The choice of the specific probability/confidence value of 90/95 is rather arbitrary. The 90/95 mix is in agreement with the "B" Basis quoted in MIL-HDBK-5C, for the characterization of material properties with test data inherently containing scatter according to a normal distribution (see Ref. 5: "At least 90 % of the population of values is expected to equal or exceed the "B" Basis mechanical property allowable, with a confidence op 95 %). Other bases quoted in reference 5 are the "A" Basis (analogous to the "B" Basis, but based on a 99/95 probability/confidence mix) and the "S" Basis where only a minimum value of the mechanical property, without any statistical assurance, is mentioned.

A fundamental base on which the specific value of 90/95 for the probability/confidence mix has been selected, is absent, however. On the contrary, it has been shown in several reliability investigations (e.g. Ref. 6) that the 90/95 mix is quite unrealistic for the in-service inspection of flaw sizes as specified in MIL-A-83444. With the NDI-techniques in the present state-of-the-art, and especially when human factors are considered, it is yet not possible to meet the stringent requirements.

With equation 6 it can be calculated how many detections n are required out of a total number N inspections of flawed specimens, to achieve a specific probability/confidence mix. For the 90/G mix with confidence level G is 50, 70, 90, 95 and 99, these values are given in table 2. It can be seen for example that for a 90/95 mix a minimum of 29 inspections have to be performed, without any misses. If one miss occurs, it should necessarily be one out of 46 inspections or more.

It is finally emphasized that the preceding discussion and calculations strictly apply to the binomial distribution of the yes-or-no detection of a present flaw. In section 3 the more general quadrinomial distribution will be discussed.

2.4 Interval methods

A considerable number of inspections is required to obtain a high degree of confidence in the estimated probability of detection for a particular flaw size. For example, it was shown in section 2.3 that when a 90/95 probability/confidence mix is required for reliable inspection, a minimum of 29 inspections without one miss has to be made. In practice, however, we will generally have a limited number of inspections for a particular flaw size, yielding low values of the lower confidence limit

p_ℓ. For example 5 detections out of 5 inspections will only yield a p_ℓ = 55 % at the required confidence level of 95 %.

Therefore, the available data points are generally grouped in intervals of flaw size to obtain a sufficient number of inspections in each interval, after which the lower confidence limit p_ℓ is calculated for all intervals. In reference 1 four methods are described for subdividing the available data points in intervals. These methods can be characterized as follows:

a) Equal-flaw-size-interval method
The total flaw-size range is divided in equal flaw-size intervals.

b) Equal-sample-size method
The total flaw-size range is divided in intervals containing the same number of data points.

c) Overlapping-interval method
The total flaw-size range is divided in overlapping intervals according to either of the two methods described before. The amount of overlap can be varied, for example a 50 % overlap for the equal-flaw-size-interval method or a 50 data points overlap for the equal-sample-size method can be chosen.

d) Optimized-probability method
The total flaw-size range is now divided in intervals in a more complicated way. The data points are arranged in order of decreasing flaw size and divided into m intervals with the largest flaws grouped in interval m. The first value of p_ℓ for the reliability curve is obtained by calculating the value of p_ℓ for each of the following intervals: m; [m+(m-1)]; [m+(m-1) + (m-2)]; ....; [m+(m-1)+(m-2)+...+1]. The maximum value of p_ℓ defined by this series of computations is plotted at the largest flaw size in interval m. Next, the data in interval m are deleted and the calculations of p_ℓ are repeated for intervals (m-1); [(m-1)+(m-2)]; ...; [(m-1)+(m-2)+...+1]. The maximum value of p_ℓ in this series is plotted at the largest flaw size in interval (m-1). These computations are repeated for decreasing m, finally resulting in m values of p_ℓ for the different intervals, and enabling the drawing of a reliability curve (Ref. 1).

In reference 7 these four different methods of presentation of reliability curves were discussed and compared. A similar study was performed in reference 8, using the inspection data of an earlier NLR-investigation (Ref. 9). It was concluded that the Optimized-Probability method is the most preferable interval method, i.e. it gives the best approximation of a monotonously increasing POD curve and it generally gives the highest numerical values for the lower confidence limit p_ℓ. In second place of presentation of NDI reliability curves comes the Overlapping-Interval method. These conclusions can be illustrated with figures 3 and 4 where the four interval methods are compared for two different inspection configurations (taken from Ref. 8).

The four interval methods described have some deficiences which are discussed in reference 10. The major deficiency is the fact that the confidence limits are greatly influenced by the specific interval method chosen. Further, in case of the overlapping of intervals, the inspection results for a particular flaw are used for more than one interval for the calculation of confidence limits. The correlation between intervals that share data, and also the influence of this correlation on the p_ℓ values in the reliability curve is unknown, however. Therefore, in reference 10 a different approach for the representation of NDI reliability is proposed, which will be discussed in section 2.5.

2.5 Regression analysis

In reference 10 a different approach (other than the binomial approach) for the representation of NDI reliability is proposed. The data set of the "Have-Cracks-Will-Travel Program" (Ref. 6), carried out for the U.S. Air Force Logistics Command by the Lockheed-Georgia company, was used to compare different functional forms of the POD curve.

In reference 6 the problem was noticed that even for equivalent inspection conditions (applied NDI method, specimen configuration, etc.), different cracks of approximately equal size did have significantly different POD's. Therefore, in reference 10 a distribution of detection probabilities was assumed at each crack size. The scatter in this distribution can be ascribed to differences in detectability due to operators, environments, and crack orientation, geometry or location. Reference 10 further introduced a density function $f_a(p)$ of the detection probabilities p for the population of specimens which have a crack length a (see Fig. 5). It was shown that the POD for a particular crack length can be seen as the mean of the detection probabilities of the individual cracks with the same crack length a:

$$POD = \int_0^1 p.f_a(p)dp. \tag{9}$$

This means that the POD(a) curve can be estimated with standard regression analysis techniques when individual estimates of the detection probabilities are available.

Next, a functional form of the POD(a) curve was searched for. Reference 10 used the "Have-Cracks" data of reference 6 to compare seven functional forms of the POD(a) curve. Beside the one used in reference 6, six other functional forms were tested by means of linear regression analysis. The different models were compared on the basis of three criteria, namely:

(1) goodness of fit, i.e. the sum of the individual deviations squared for the different data points is a minimum (deviation is the difference between the real probability and the value as determined from the assumed curve). The curve having this property is also called a least-squares curve;

(2) normality of deviations from fit (with the Shapiro-Wilks W test, Ref. 17);

(3) equality of variance of deviations from fit for all crack lengths (with the Bartlett's test, Ref. 17).

It was concluded in reference 10 that the log odds-log scale model among the seven investigated models provided the best fit to the POD(a) function for the "Have-Cracks" data.

The log odds-log scale model makes use of the following regression equation of y on x:

$$y = A + Bx \quad \text{where } x, y = \text{variable} \qquad A, B = \text{regression constants} \tag{10}$$

and applies the transformations:

$$y = \ln\{p/(1-p)\} \quad ; \; p = \text{probability of detection}$$

$$x = \ln(a) \qquad a = \text{crack length.} \tag{11}$$

These transformations result in the following equation for the probability of detection:

$$p = POD(a) = \frac{\exp(A+B.\ln a)}{1 + \exp(A+B.\ln a)}. \tag{12}$$

The constants A and B can be calculated with a standard linear regression analysis using the available data of (a,p). A description of this analysis can be found in any textbook on statistical theory; e.g. in reference 2, chapter 18 or reference 4, chapter 9.6.3.
With the equations:

$$\begin{aligned} y &= A + B.x \\ \Sigma y &= N.A + B.\Sigma x \\ \Sigma xy &= A.\Sigma x + B.\Sigma x^2 \end{aligned} \tag{13}$$

we can solve for A and B:

$$A = \frac{(\Sigma y).(\Sigma x^2) - (\Sigma x).(\Sigma xy)}{N\Sigma x^2 - (\Sigma x)^2}$$

$$B = \frac{N.\Sigma xy - (\Sigma x).(\Sigma y)}{N.\Sigma x^2 - (\Sigma x)^2}. \tag{14}$$

Lower confidence limits on the POD(a) function can also be calculated. Since the log-odds transformation of eq. 11 is monotonic, the inverse transformation of the confidence limit on a mean y(a) will yield the confidence limit on p(a).

Further, the principle of maximum likelihood is used in reference 10, to calculate the maximum likelihood estimates $\hat{A}$ and $\hat{B}$ of the parameters A and B in equation (10). Most important is that grouping of data is not required but that instead the observed outcomes of the yes-or-no detection of a particular flaw are being used. This approach is especially valuable for experiments in which each flaw is inspected only once. The estimates $\hat{A}$ and $\hat{B}$ can be calculated with the following equations (procedure adopted from Ref. 11):

$$0 = \sum_{i=1}^{N} Z_i - \sum_{i=1}^{N} \frac{\exp(\hat{A}+\hat{B}.\ln(a_i))}{1 + \exp(\hat{A}+\hat{B}.\ln(a_i))}$$

$$0 = \sum_{i=1}^{N} Z_i \ln(a_i) - \sum_{i=1}^{N} \frac{\ln(a_i).\exp(\hat{A}+\hat{B}.\ln(a_i))}{1 + \exp(\hat{A}+\hat{B}.\ln(a_i))} \qquad (15)$$

where $Z_i = 1$ if the flaw is detected and $Z_i = 0$ if it is not. The maximum likelihood estimate of POD(a) is then obtained by substitution of the estimates $\hat{A}$ and $\hat{B}$ in eq. 12. In reference 10 further formulae are given for the calculation of lower confidence limits on POD(a). An example of the use of maximum likelihood estimates is given in figure 6, where the plotting of the lower 95 % confidence limit from the maximum likelihood analysis is compared with the optimized probability method (section 2.4), using the inspection results of 361 cracks.

It can be concluded that the different model approach in reference 10 for the representation of NDI reliability is most promising. The log odds-log scale model, by making use of linear regression analysis, has shown to be a more adequate representation of the POD function (closer to the true POD, and less scatter in the distribution of the estimates) than the binomial approach. NDI experiments in which each flaw is inspected only once can be treated with maximum likelihood analysis. For both the linear regression analysis and the maximum likelihood analysis, lower confidence limits on the POD function can be calculated.

3 QUADRINOMIAL DISTRIBUTION

3.1 Treatment of false calls

In the preceding chapter the attention was focused on the probability of detection (POD) of flaws in flawed specimens. The inspection could therefore be described by means of a binomial distribution. In practice, however, also specimens without flaws are inspected. These specimens give rise to a second distribution, namely the yes-or-no detection of a flaw in unflawed specimens. Therefore, the result of a nondestructive inspection is more correctly described by means of a quadrinomial distribution (Fig. 7). The fraction of actually flawed specimens can be divided into a fraction of detected flaws (correct rejection) and a fraction of missed flaws (false acceptance). The fraction of unflawed specimens can, in turn, be divided into a fraction of specimens recognized as unflawed (correct acceptance) and a fraction of rejected but actually unflawed specimens (false rejection). In analogy with the probability of detection (POD) for the flawed specimens we can now define a probability of recognition (POR) for the unflawed specimens.

Both POD and POR are essential inspection characteristics. Their relative importance depends upon the criteria placed upon the inspection performance. From a safety point of view POD is the essential characteristic, while POR is more important from an economics point of view. It seems reasonable to include both POD and POR, each with specified weighting factors, in a characterization of the inspection performance. In the next section different methods will be discussed to obtain a general measure for the characterization of the reliability of a NDI technique.

3.2 Measure of reliability

The four possible outcomes of an inspection (quadrinomial distribution) can be arranged in a fourfold matrix. In reference 12 the composition of such a fourfold (2x2) contingency table is described, given an inspection configuration of n specimens of which c_1 are flawed (see Fig. 7). The n specimens are divided into a set F of c_1 actually flawed specimens and the complimentary set $\tilde{F}$ of c_2 actually unflawed specimens. Upon inspection the specimens of set F will yield a fraction a_{11} of correct rejection and a fraction a_{21} of false acceptance ($c_1=a_{11}+a_{21}$). The specimens of set $\tilde{F}$ will yield a fraction a_{12} of false rejection and a fraction a_{22} of correct acceptance ($c_2=a_{12}+a_{22}$). Next to the sets F and $\tilde{F}$ the inspection will result in a set M of r_1 specimens suspected of being flawed and the complimentary set $\tilde{M}$ of r_2 specimens marked as unflawed. M consists of the fractions a_{11} and a_{12} with $r_1 = a_{11} + a_{12}$, $\tilde{M}$ consists of the fractions a_{21} and a_{22} with $r_2 = a_{21} + a_{22}$. Evidently the total number of specimens n is equal to (r_1+r_2) or (c_1+c_2). The foregoing is illustrated in detail in figure 7.

Independence of effects can be tested by means of standard chi-square calculation. The test statistic is:

$$\chi^2 = \frac{n.\Delta^2}{r_1 r_2 c_1 c_2} \qquad (16)$$

where Δ is the determinant of the fourfold matrix.
χ^2 can be written in an alternative way as follows:

$$\chi^2 = n. \frac{\Delta}{r_1 r_2} \cdot \frac{\Delta}{c_1 c_2} = n.\left(\frac{a_{11}}{r_1} - \frac{a_{21}}{r_2}\right) \cdot \left(\frac{a_{11}}{c_1} - \frac{a_{12}}{c_2}\right)$$

This test statistic is a general measure for the ability of an inspector to distinguish flaws better than from random marking: high inspection performance will result in a high value of the test statistic. Although the chi-square test is acceptable for inspector ranking purposes it is pointed out in reference 12 that the mean square contingency $\varphi = \sqrt{\chi^2/n}$ is a more appropriate measure; mainly because of the influence on χ^2 of the size n of the data base.

For a more general measure of the inspection performance - or reliability of inspection - φ still has some drawbacks, for example the difficulty of how to interpret φ in probablistic terms. In reference 12 several alternatives were discussed; it was concluded that the Somers's d coefficient (Ref. 13), defined as

$$d = \frac{a_{11}a_{22} - a_{12}a_{21}}{c_1 c_2} \tag{17}$$

is probably a more suitable measure.
Somers's d, rewritten as:

$$d = \frac{a_{11}}{c_1} - \frac{a_{12}}{c_2} \tag{18}$$

can be interpreted as the difference in conditional probabilities of a detected flaw versus a false call.

We are still seeking for a more general measure for the characterization of NDI reliability. We therefore rewrite Somers's d as follows:

$$d' = d + 1 = \frac{a_{11}}{c_1} + \frac{a_{22}}{c_2} = POD + POR \tag{19}$$

which is nothing else but the sum of the conditional probabilities of correct rejection and correct acceptance.

In section 2.2 the statistical variation in the estimated POD (point estimate $\bar{p}$) was discussed, which resulted in the calculation of lower bound values of the POD at a specified confidence level G. In analogy with this we can calculate a lower bound value of the POR at the same (or another) confidence level G. We now define an adjusted Somers's d" as follows:

$$d'' = POD_\ell + POR_\ell \qquad (20)$$

where the subscript ℓ denotes the lower bound value of the POD and POR, respectively.

The contribution of POD_ℓ and POR_ℓ to d" need not necessarily be equal, depending on which value is set on these factors by considerations of safety and economics. A more general measure thus is

$$d'' = \alpha . POD_\ell + (1-\alpha) . POR_\ell \qquad (21)$$

with α a specified weighting factor.

The construction of d" as a general measure for the characterization of NDI reliability has one major drawback, namely the fact that different combinations of POD_ℓ and POR_ℓ can give the same value of d". It is therefore doubtful whether a single parameter will give an appropriate measure of NDI reliability. We suggest instead to keep separate the different conditional probabilities, and to specify a minimum value for both the POD and POR at a specified confidence level G. We finally associate reliability of NDI with the ability of an inspector to perform a nondestructive inspection, under specified inspection conditions and procedures, with a certain probability of detection (POD) for flaws of a particular size and with a certain probability of recognition (POR) for the unflawed parts, both at a specified confidence level G.

3.3 Relative operating characteristic

In reference 14 a model was presented to provide a measure of the inherent accuracy with which a technician, technique, or total system can discriminate flawed from unflawed materials. It was concluded that one separate index, derived from modern theory of signal detection, is able to determine the balance between the proportions of true and false detections observed at a given time.

Basis of the approach in reference 14 is the plotting of the probability of true finds (a_{11}/c_1) against the probability of false calls (a_{12}/c_2), yielding the so-called "relative operating characteristic" or ROC. In figures 8 and 9 examples of this procedure are given for two inspection cases; the inspections were performed by about 100 USAF technicians of different proficiency. Clearly, a definite relationship between the percentages of true finds and false calls can be observed at each degree of inspector proficiency: a higher percentage of true finds can be reached at the cost of an increase in the number of false calls. This shift is merely the result of the use of a different decision criterion at a same proficiency level.

To provide an accuracy index of the technician proficiency, several possibilities to index the location of a ROC were discussed in reference 14. It was concluded that the most appropriate measure is the proportion of area in the unit square that is located beneath the ROC. As index of the decision criterion, determining a particular point on the ROC; the slope of the line tangent to the ROC curve at that point, was recommended.

Also other models and indices of accuracy were discussed in reference 14, among which the mean-square contingency φ and Somers's d. Mainly because of the discrepancy between the ROC curves derived theoretically from these models and ROC curves from typical data (linear instead of curved lines), the other models were regarded as incorrect.

In our opinion, however, it will be extremely difficult in NDI practice to produce sufficient points of a ROC for a reliable calculation of the accuracy index (defined here as the proportion of area beneath the linear ROC that is fitted to the data points). This namely implies that, either one must have the availability of inspections from inspectors who handled very different inspection criteria, or that an inspector deliberately has to judge the inspection result against different decision criteria. Especially with NDI where still many inspections are performed in qualitative way (judgement of penetrant or magnetic particle indications, radiograph, eddy-current CRT screen, etc.) this will not be an easy task; in other words, it will be difficult to "reliably" determine the accuracy index of NDI, as defined in reference 14.

4 RELIABILITY DEMONSTRATION PROGRAM

To determine the reliability of NDI techniques, often a so-called "reliability demonstration program" is conducted. This generally means the inspection of a large number of identical specimens after which data verification takes place by means of a fractographic analysis. The most extensive program performed until now is known as the "Have Cracks-Will Travel" program by the Lockheed-Georgia Company for the U.S. Air Force Logistics Command (Ref. 6). In this program approximately 300 Air Force technicians performed ultrasonic, eddy-current, penetrant, and radiographic inspections on actual aircraft structural samples containing fatigue damage. A total of about 22.000 inspections were made on 174 cracks. It was concluded from this program that the reliability of most nondestructive inspection methods is less than had been assumed; the often quoted 90/95 probability/confidence mix could almost never be attained for flaw sizes as defined in MIL-A-83444. Also it was concluded that large differences in the inspection results occur, and that these differences are primarily caused by human factors. A drawback of this program, however, is that the treatment of false calls was not included.

To provide a common baseline for demonstrating flaw detection capabilities of NDI methods, the American Society for Nondestructive Testing (ASNT) prepared a guideline for the preparation of specific reliability demonstration programs (Ref. 15). This guideline includes e.g. the personnel and environment variables, specimen preparation, parameters of the different NDI methods, and statistical treatment of the inspection results.

Although the document in reference 15 specifies that a sufficient number of unflawed specimens should be randomly mixed with the flawed specimens to provide a quadrinomial distribution of the inspection results, only attention is focused on the probability of detection of the flaws in the flawed specimens while ignoring the treatment of false calls. Reliability is also defined in the document as "the probability of detecting a crack in a given size group under the inspection conditions and procedures specified in the inspection procedure document". In our opinion, however, the treatment of false calls may not be ignored in a reliability demonstration program. As explained in sections 3.1 and 3.2 where in analogy with the probability of detection (POD) a probability of recognition (POR) was defined, both POD and POR are essential inspection characteristics and should be included in a general assessment of the reliability of an inspection technique.

If the treatment of false calls is included in a demonstration program there still is the drawback of the different inspection conditions for the inspectors. These conditions will not entirely resemble the physical and mental conditions under which the inspectors perform in their every day field environment. This could have a considerable influence on the results of the demonstration program. For example if only highly motivated inspectors are included in the program, this could result in a reliably detectable flaw size which is unlikely small.

In summary, it can be said that the human factor is still the main drawback in the execution of a reliability demonstration program. There is no denying that the inspection results will show a large degree of scatter, which limits for example the demonstration of a certain flaw detection capability. The only way to tackle this problem of non-reproducibility is to collect a large amount of inspection results by as many inspectors as possible, using different inspection e.g. specimen configurations, and to analyse the inspection results with the methods discussed in section 2.5 e.g. with regression analysis. The results should finally be expressed in terms of the probability of detection for flaws of a particular size as well as the probability of recognition for the unflawed parts, both at a specified confidence level.

5 CONCLUSIONS

1. The Optimized-Probability method is the most preferable interval method for subdividing available data points in intervals, i.e. it gives the best approximation of a monotonously increasing POD curve and it generally gives the highest numerical values for the lower confidence limit p_{ℓ}.

2. The choice of the specific probability/confidence mix of 90/95 for reliable inspection is rather arbitrary and not based on current detection capabilities of most nondestructive inspection methods.

3. The log odds-log scale model, by making use of linear regression analysis, has shown to be a more adequate representation of the probability of detection function than the binomial approach.

4. The concept of reliability should be discussed in terms of the quadrinomial distribution of the possible outcomes of NDI: not only the probability of detection (POD) of flaws in flawed specimens, but also the probability of recognition (POR) for the unflawed specimens is an essential inspection characteristic.

5. It is doubtful whether a single parameter will give an appropriate measure of NDI reliability; instead it is recommended to establish minimum values for both the POD and POR at a specified confidence level.

6. A reliability demonstration program should not only consider the POD of flaws in flawed specimens, but also the treatment of false calls; also then it will be extremely difficult, because of different human factors, to establish results which are generally acceptable in practice.

6 REFERENCES

1. Packman, P.F., et al. — Reliability of flaw detection by nondestructive inspection.
ASM Metals Handbook, Vol. 11, 8th Edition
Metals Park, pp. 414-424,
Ohio, 1976

2. Hald, A. — Statistical theory with engineering applications.
University of Copenhagen, Published by
John Wiley & Sons, Inc.,
Copenhagen, August 1951

3.	United States Air Force	Military Specification MIL-A-83444, Airplane damage tolerance requirements, July 1974
4.	Wood, H.A., Engle, R.M.	USAF Damage tolerant design handbook: Guidelines for the analysis and design of damage tolerant aircraft structures. AFFDL-TR-79-3021, Wright-Patterson AFB, Ohio, March 1979
5.		MIL-HDBK-SC Military standardization handbook, Metallic materials and elements for aerospace vehicle structures, September 1976
6.	Lewis, W.H., Dodd, B.D., Sproat, W.H., Hamilton, J.M.	Reliability of nondestructive inspections. Final report, Report no. SA-ALC/MME76-6-38-1, United States Air Force, San Antonio, Air Logistics Center, Kelly AFB, Texas, December 1978
7.	Yee, B.G.W., Chang, F.H., Coughman, J.C., Lemon, G.H., Packman, P.F.	Assessment of NDE reliability data. NASA CR-134991. National Aeronautics and Space Administration, Lewis Research Center, Cleveland, Ohio, 1976
8.	De Graaf, E.A.B.	Presentation of NDT-reliability curves, Volume 1 and 2. NLR TR 82076 L, Amsterdam, 1982
9.	De Graaf, E.A.B., De Rijk, P.	Evaluation of non-destructive inspection methods applied to milled step-down areas. NLR TR 78110 U, Amsterdam, 1978

10. Berens, A.P., Hovey, P.W. — Evaluation of NDE reliability characterization. AFWAL-TR-81-4160, University of Dayton Research Institute, December 1981

11. Cox, D.R. — The analysis of binary data. Methuen and Co., Ltd., London, 1970

12. Sharp, H., Sproat, W.H. — Treatment of false calls in evaluating nondestructive inspection proficiency. Journal of nondestructive evaluation, Volume 2, nos. 3/4, 1981

13. Somers, R.H. — A new asymmetric measure of association for ordinal variables, Amer. Sociol. Rev. 27, pp. 799-811, 1962

14. Swets, J.A. — Assessment of NDT systems. Materials Evaluation, Volume 41, no. 11, pp. 1294-1303, October 1983

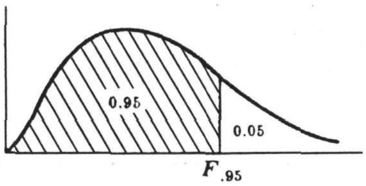
15. Rummel, W.D., ASNT — Recommended practice for a demonstration of nondestructive evaluation (NDE) reliability on aircraft production parts. Materials Evaluation, Volume 40, no. 9, p. 922, August 1982

16. Pearson, E.S., Hartley, H.O. — Biometrika tables for statisticians, Volume 2, Table 5, page 178, 1972

17. Anderson, V.L., McLean, R.A. — Design of experiments, Marcel Dekker, Inc. New York, 1974

TABLE 1

95th Percentile Values (0.05 Levels), F_{95}, for the F Distribution



ν_1 degrees of freedom in numerator

ν_2 degrees of freedom in denominator

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(Table 5 from reference 16, page 178)

TABLE 2

Number of inspections and permitted failures to demonstrate a probability of detection of 90 % or more at indicated confidence level

NUMBER OF FAILURES	PROBABILITY OF DETECTION 90 % CONFIDENCE LEVEL [%]				
	50	70	90	95	99
0	7	12	22	29	51
1	17	24	38	46	72
2	27	36	52	61	90
3	37	47	65	76	106
4	47	58	78	89	122
5	57	70	91	103	137
6	67	81	104	116	152
7	77	91	116	129	167
8	87	102	128	142	181
9	97	113	140	154	195
10	107	124	152	167	209

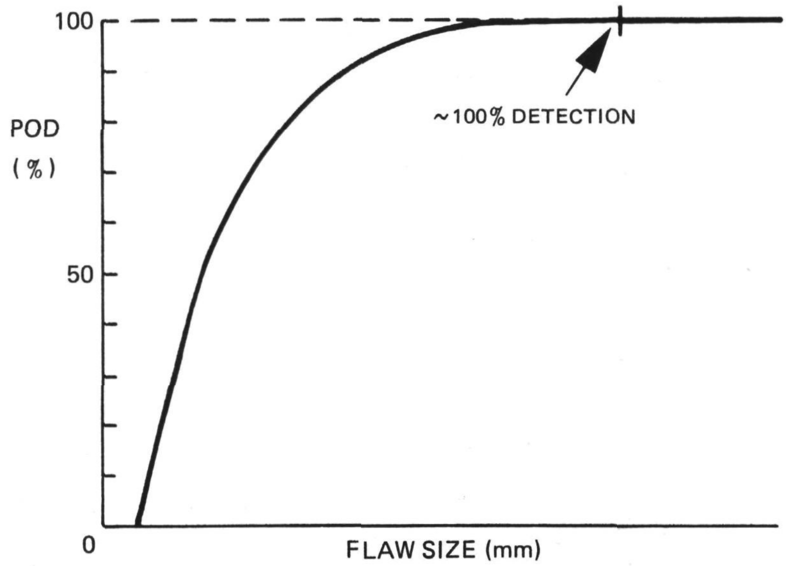


Fig. 1 Probability of detection (POD) curve

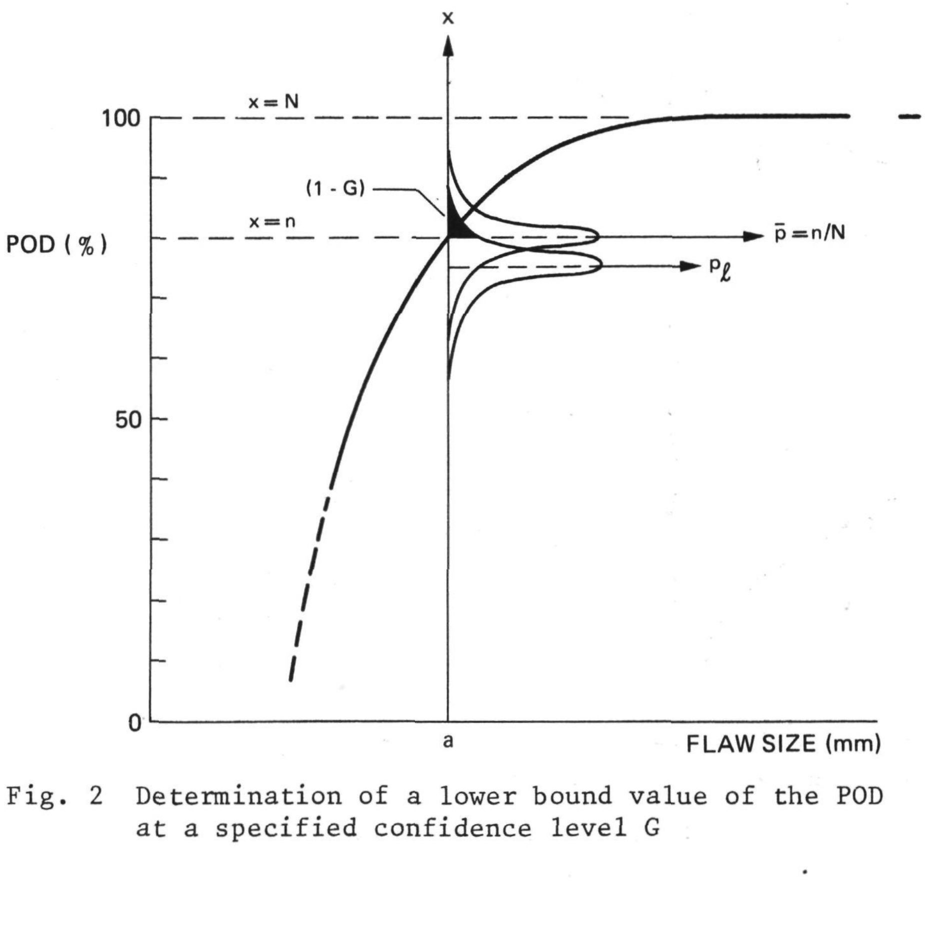


Fig. 2 Determination of a lower bound value of the POD at a specified confidence level G

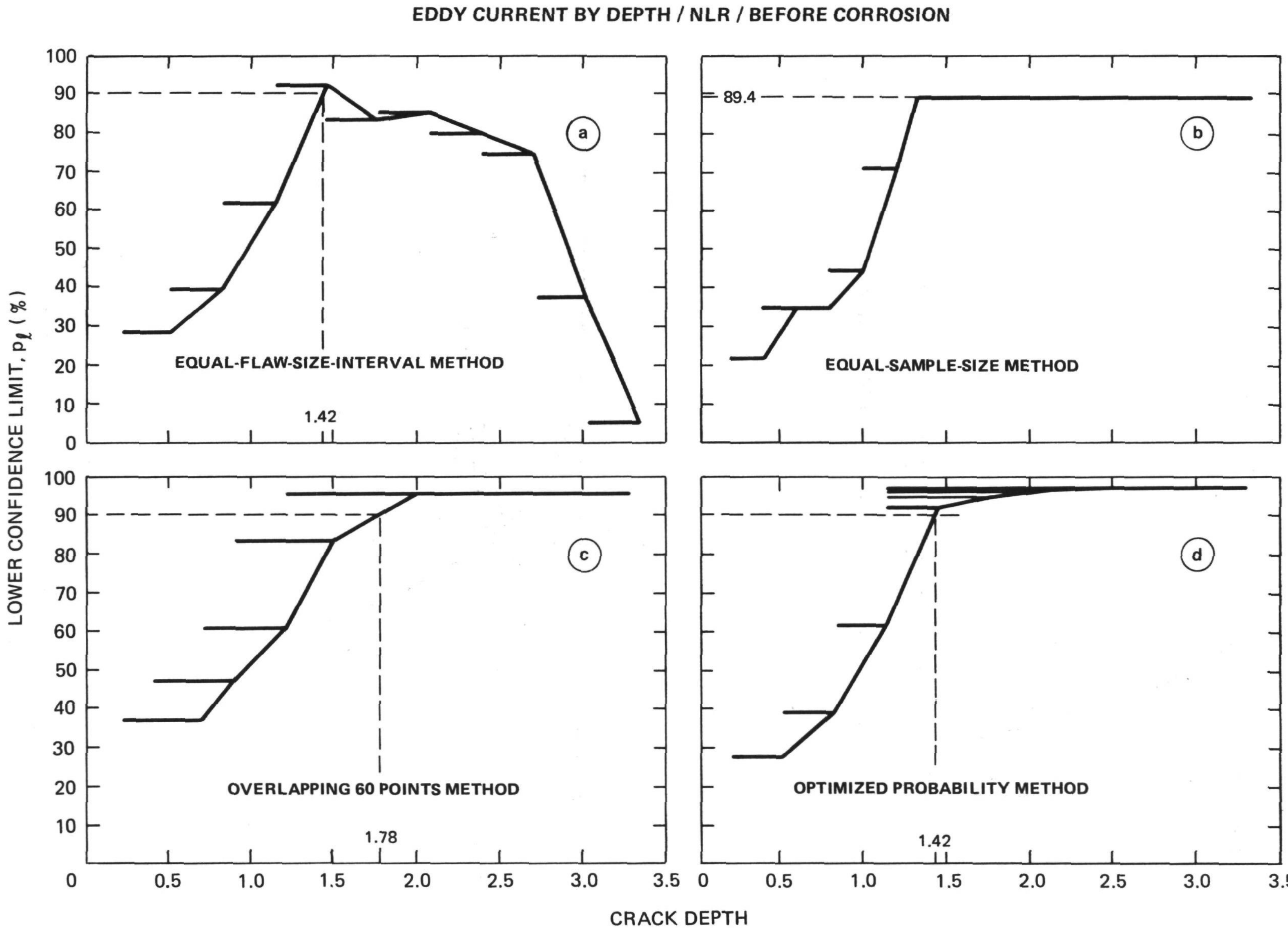


Fig. 3 Reliability curves, charts of lower confidence limits, p_ℓ, of fatigue cracks of various sizes in specimens of reference 9. Horizontal bars denote the applicable flaw size interval. To produce conservative graphs the calculated data were considered to be applicable to the largest size in each interval; therefore the right hand sides of the horizontal lines were connected (Fig. 7 from reference 8)

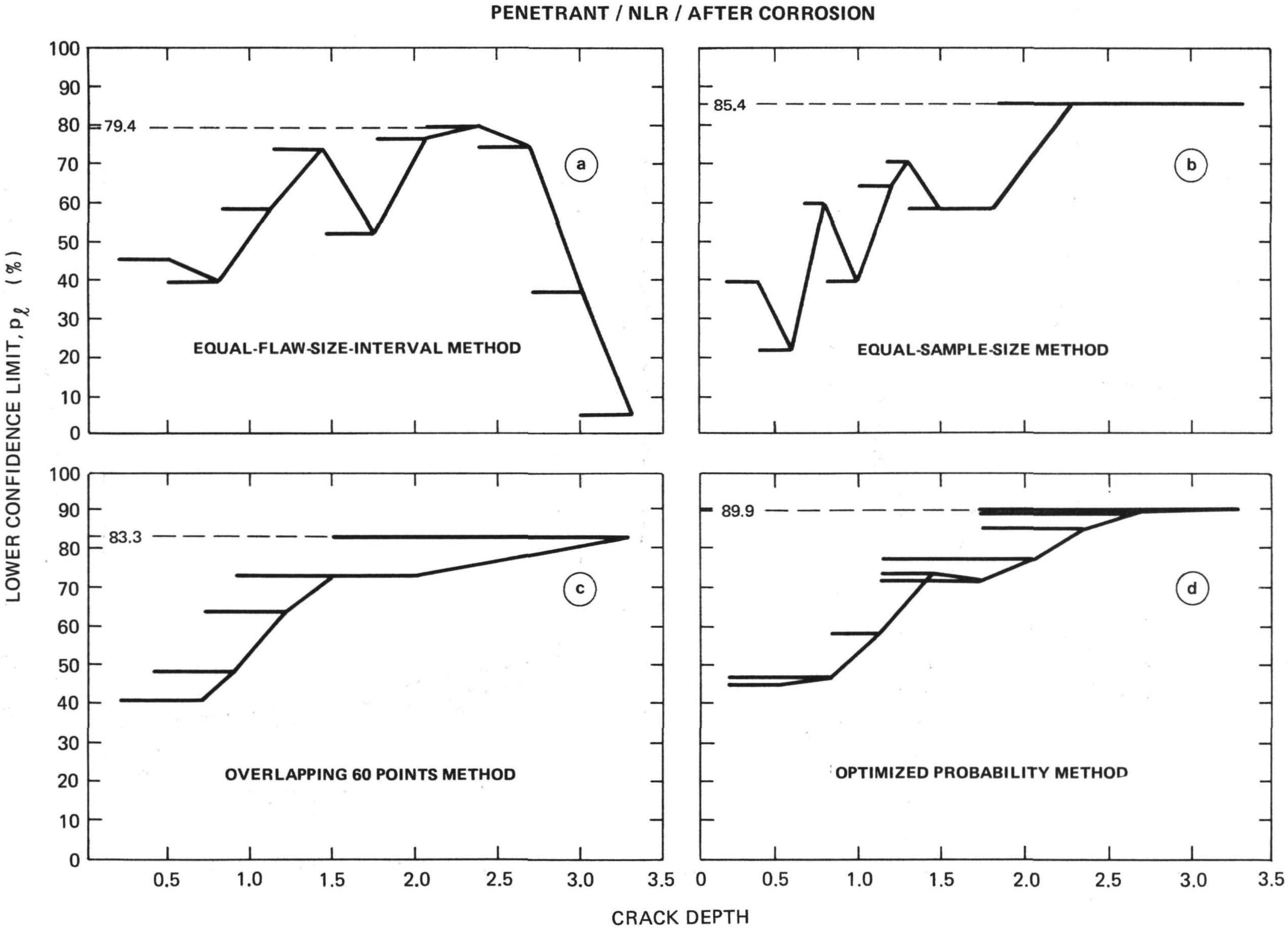


Fig. 4 Ditto (Fig. 8 from reference 8)

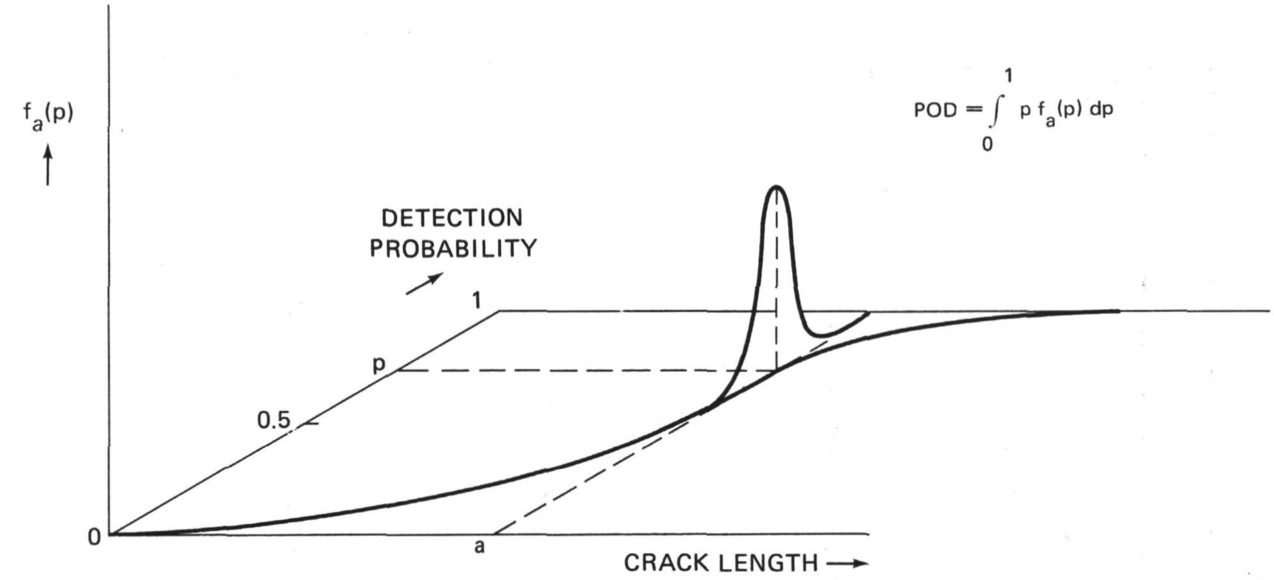


Fig. 5 Representation of a distribution of detection probabilities $f_a(p)$ at a crack length a

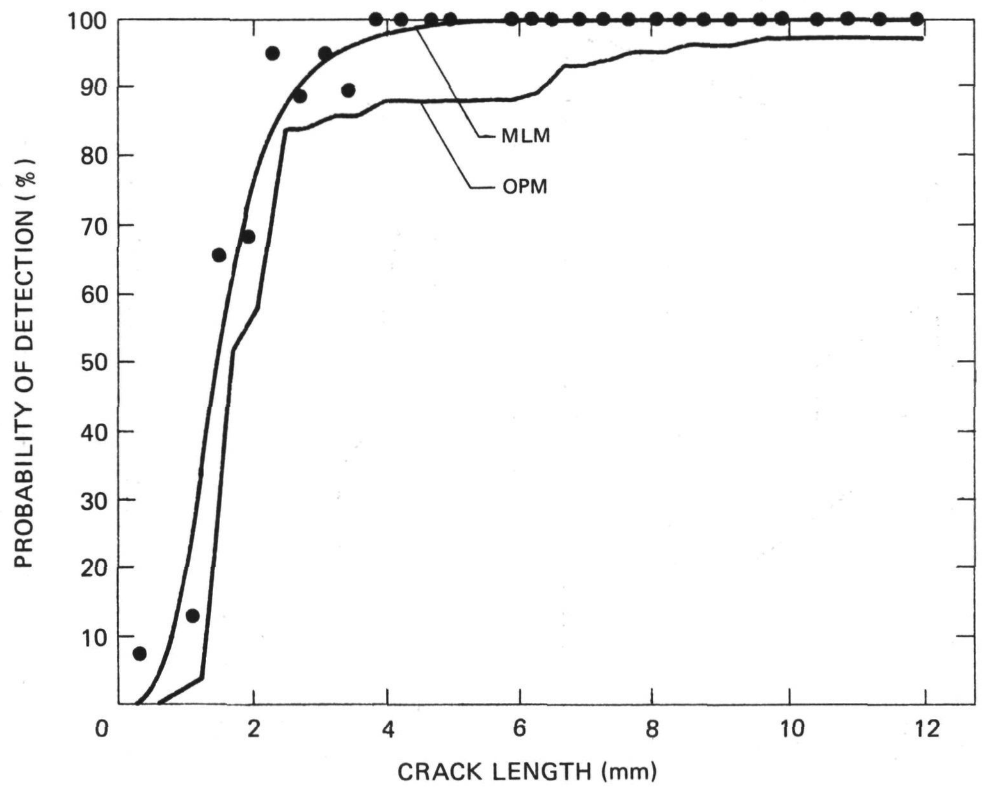
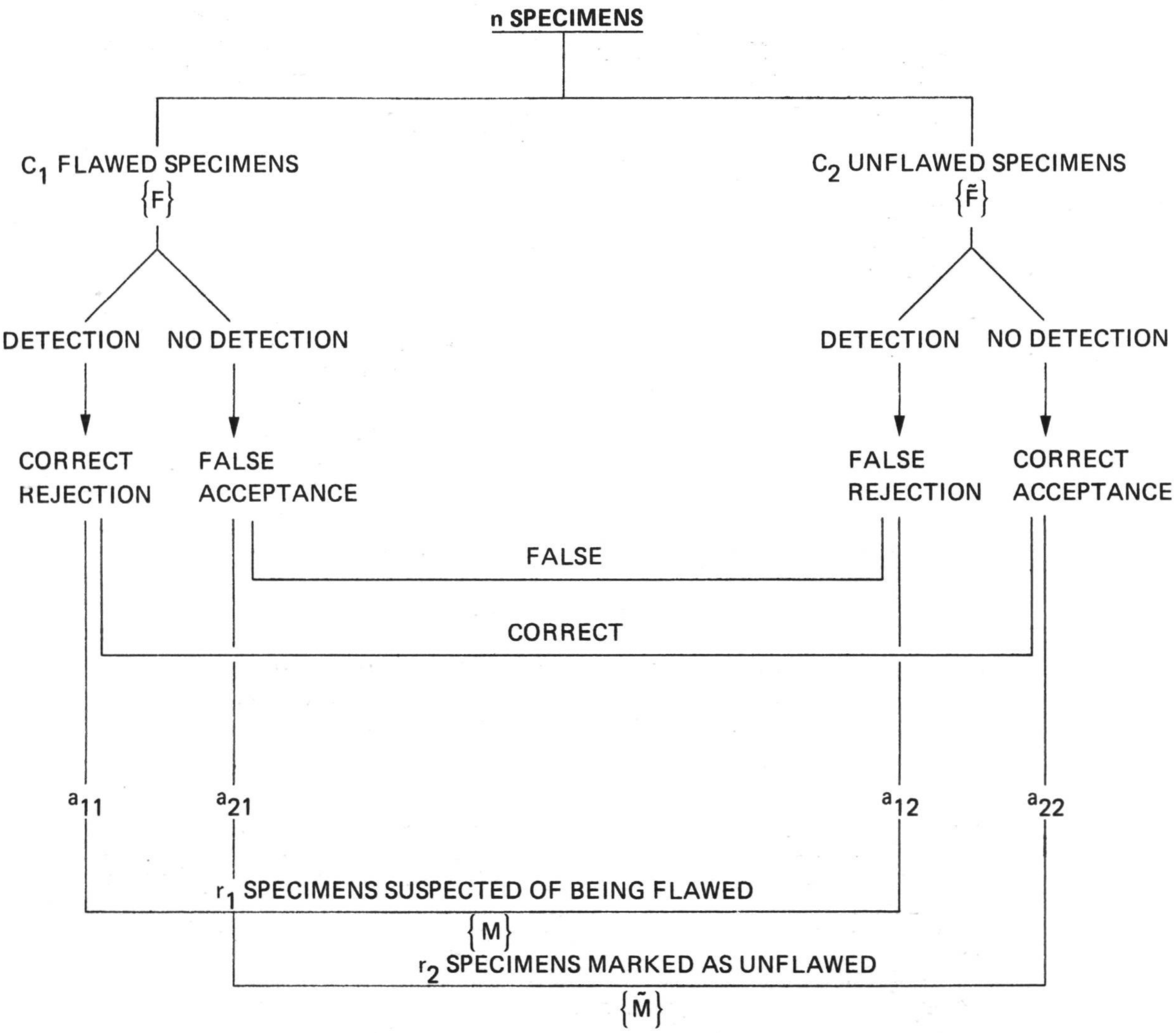


Fig. 6 Comparison of Confidence Bounds Obtained with the Maximum Likelihood Method and the Optimized Probability Method (Fig. 9 from reference 10)



	F	$\tilde{F}$	
M	a_{11}	a_{12}	r_1
$\tilde{M}$	a_{21}	a_{22}	r_2
	c_1	c_2	n

$$n = r_1 + r_2 = c_1 + c_2$$

Fig. 7 The four possible outcomes of an inspection -quadrinominal distribution-

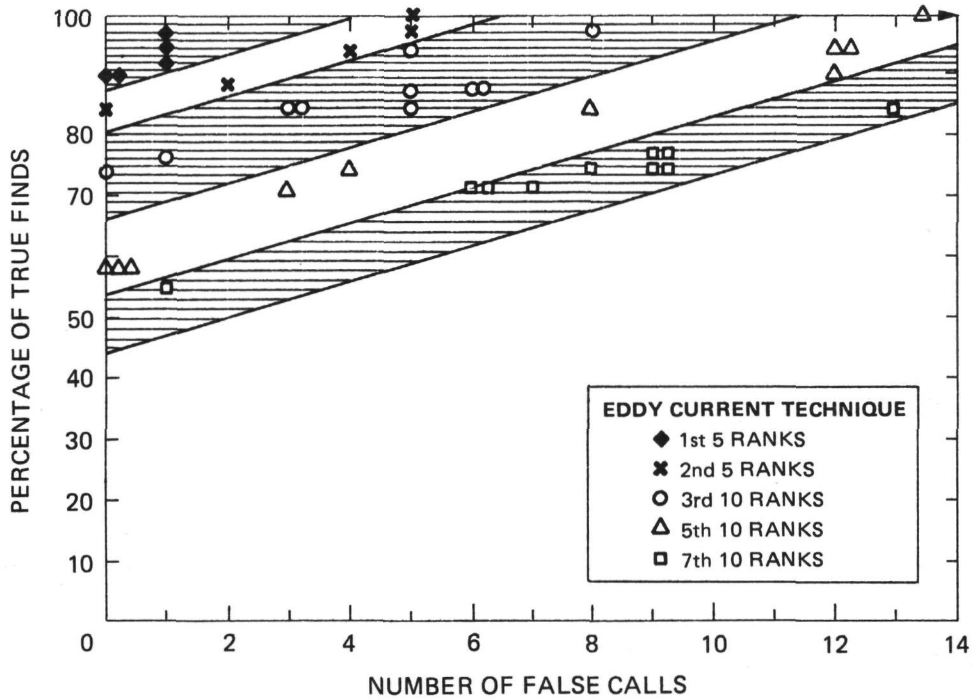


Fig. 8 Percentage of true finds versus number of false calls for selected groups of NDT technicians ranked according to proficiency. Data from an eddy current technique

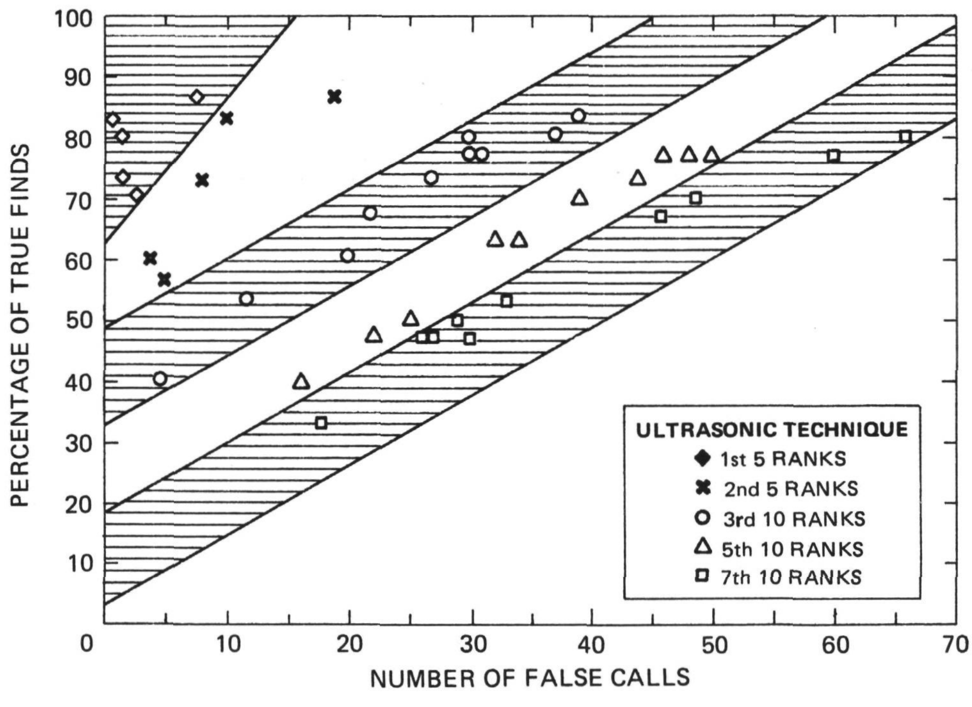


Fig. 9 Same graph as figure 8. Data from an ultrasonic technique. (Figure 1 from reference 14)

NATIONAAL LUCHT- EN RUIMTEVAARTLABORATORIUM

Anthony Fokkerweg 2, 1059 CM AMSTERDAM
Postbus 90502 , 1006 BM AMSTERDAM
Tel. (020) 511313 - telex 11118 (nlraa nl)
Telegram-adres: Windtunnel Amsterdam