REFERENCES

TECHNOMETRICS, FEBRUARY 1994, VOL. 36, NO. 1

A *D*-Optimality-Based Sensitivity Test

Barry T. NEYER

EG&G Mound Applied Technologies
Miamisburg, OH 45343-3000

Sensitivity tests are often used to estimate the parameters associated with latent continuous variables that cannot be measured. For example, each explosive specimen has a threshold. The specimen will detonate if and only if an applied shock exceeds this value. Since there is no way to determine the threshold of an individual, specimens are tested at various levels to determine parameters of the population. A new test described here produces efficient estimates of the parameters of the distribution, even with limited prior knowledge. This test efficiently characterizes the entire distribution and desired percentiles of any population.

KEY WORDS: Bruceton method; Langlie method; Optimal design; Probit method.

Sensitivity tests are often used to estimate the parameters associated with latent continuous variables that cannot be measured. For example, in testing the sensitivity of explosives to shock, each specimen is assumed to have a critical stress level or threshold. Shocks larger than this level will always explode the specimen, but smaller shocks will not lead to explosion. Repeated testing of any one sample is not possible because the stress that is not sufficient to cause explosion nevertheless will generally damage the specimen. To measure probability of response, samples are tested at various stress levels and the response or lack thereof is noted.

Explosives designers are often interested in determining the *all-fire* level, usually defined as the level of shock necessary to cause 99.9% of the specimens to fire. (Some designers seek the 99.99% or 99.9999% levels.) A distribution-independent method of estimating a level would require *at least* one response and one nonresponse at the specified level. Since sample sizes are usually far smaller than the many thousands required to estimate these extreme levels, the explosive designer generally relies on parametric methods. Parametric methods also allow the experimenter to characterize the population as a whole and to evaluate process variation. Parametric designs test specimens at several stress levels. The parameters are estimated by maximum likelihood or other techniques.

This article describes a new sensitivity test based on a known probability response curve. It has advantages over many previously described sensitivity tests, especially if the parameters of the probability distribution are not well known in advance. This procedure has a starting algorithm that quickly produces unique estimates of the parameters, regardless of how close the parameters of the population are to the initial guesses. It uses a design motivated by *D*-optimality considerations for the remaining samples to maximize knowledge of the parameters of the curve.

A *c*-optimal design would allow more precise estimation of *one* quantile by concentrating the tests near the specified level. The estimate would also be independent of the distribution. The *c*-optimal design, however, does not efficiently provide knowledge of the form of the whole population. (A *c*-optimal design for estimating quantiles in the tails of a distribution would concentrate tests at two points in the distribution. See Wu [1988] for further details.) A *D*-optimal design provides efficient estimates of the parameters of the distribution. It allows relatively efficient determination of *all* quantiles of the population, but the estimates are distribution dependent.

A *D*-optimal design in many cases could be of more use to engineers than a *c*-optimal design, even when the engineer is only interested in one extreme quantile. Suppose that an engineer wants to test new explosive mixtures for greater sensitivity. A *c*-optimal design should allow the engineer to determine the all-fire level more efficiently than a *D*-optimal design. The engineer could then restrict future study to mixtures with smaller thresholds. Knowledge of the entire response curve, however, would allow the experimenter to further investigate potentially promising mixtures. For example, a mixture with a larger all-fire level but a smaller scale parameter might give insight into methods of improving process control. And a mixture with lower mean but larger scale parameter might yield a lower all-fire level if the process could be brought under better control. In addition, it is often easier to detect significant changes in the location and scale parameters than in an extreme quantile.

Section 1 of this article presents a brief review of the theory, Section 2 explains the new sensitivity test and illustrates it with an example, Section 3 discusses

the Monte Carlo comparison of the various tests, Section 4 lists variations on this new test, and Section 5 discusses several practical considerations for sensitivity testing.

1. ESTIMATES OF μ AND σ

The rest of this article assumes that the probability response function takes the form $P[(x - \mu)/\sigma]$, where the x's are the (transformed) stimulus levels and μ and σ are the location and scale parameters. The examples assume that P is a normal distribution function. If the probability function is not normal, then often a suitable transformation of the stress levels will make the distribution normal. (For many fields, x is the log of the applied stress.) The exact form of the distribution curve will be stated where it makes a difference to the discussion.

Maximum likelihood estimates (MLE's) of the parameters are used because they are relatively easy to compute and have desirable asymptotic properties. The use of the likelihood function in the analysis of sensitivity tests has been treated previously (Cornfield and Mantel 1950; Golub and Grubbs 1956). A limited review follows.

Let x_i be the (transformed) stimulus level for the ith test, n_i be the number tested at this level, and p_i be the proportion of samples that responded. (The responses are often called successes and the nonresponses failures.) Let $P[(x_i - \mu)/\sigma]$ be a known distribution function with μ and σ the unknown parameters. Define $z_i = (x_i - \mu)/\sigma$, $Q(z) = 1 - P(z)$, and $q_i = 1 - p_i$. The likelihood function, $L(\mu, \sigma)$, is the probability of obtaining the given test results with the specified μ and σ. It is given by

$$L(\mu, \sigma) = \prod_i \binom{n_i}{p_i n_i} P(z_i)^{n_i p_i} Q(z_i)^{n_i q_i}. \quad (1)$$

The values, $\hat{\mu}$ and $\hat{\sigma}$, which maximize the likelihood function, are the MLE's. Unique MLE's will be obtained if the successes and failures overlap; that is, the smallest success is smaller than the largest failure (Silvapulle 1981).

The Fisher information matrix (Kendall and Stuart 1967) provides a measure of the information on the parameters of the distribution from the test data. It is obtained by computing expected values of second derivatives of the log of the likelihood function. The information matrix is given by

$$I_{jk} = E\left(\frac{1}{L}\frac{\partial L}{\partial \theta_j} \cdot \frac{1}{L}\frac{\partial L}{\partial \theta_k}\right). \quad (2)$$

For sensitivity tests, $E(p_i) = P(z_i)$. Let $\theta_0 = \mu$ and $\theta_1 = \sigma$, and define

$$J_j(z_i) = P'^2(z_i) z_i^j / [P(z_i)Q(z_i)\sigma^2]. \quad (3)$$

The information matrix for sensitivity tests has the elements

$$I_{jk} = \sum_i n_i J_{j+k}(z_i), \quad (4)$$

found by adding the $J_j(z_i)$ functions evaluated for each test level.

The asymptotic variances of the MLE's are given by terms in the inverse of the information matrix:

$$\text{var } \hat{\mu} = I_{11}/(I_{00}I_{11} - I_{01}^2) \quad (5)$$

and

$$\text{var } \hat{\sigma} = I_{00}/(I_{00}I_{11} - I_{01}^2). \quad (6)$$

Figure 1 shows the three functions $J_j(z)$ as a function of the normalized stimulus level, z, for normal probability. Banerjee (1980) obtained similar results. Similar curves can be obtained for other probability functions.

Since $P'(z)$ is symmetric, the functions $J_0(z)$ and $J_2(z)$ are symmetric functions of z, but $J_1(z)$ is an antisymmetric function. For large sample sizes, $I_{01} \ll I_{00}, I_{11}$ if the test levels are chosen symmetrically. Under this condition, minimization of the asymptotic variances for $\hat{\mu}$ ($\hat{\sigma}$) is equivalent to maximization of the value I_{00} (I_{11}). (It is known in optimal-design theory that optimal designs can always be made symmetric for symmetric probability density functions.)

Figure 1 shows that the asymptotic variance for $\hat{\mu}$ will be minimized if the tests are concentrated near

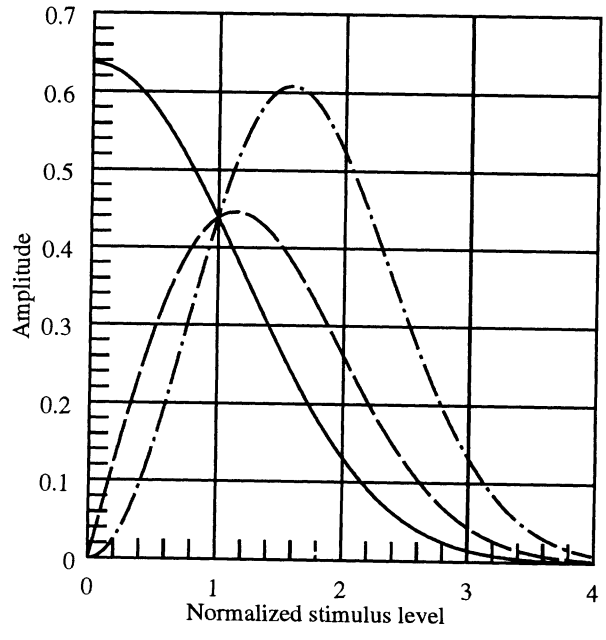

Figure 1. Normal Sensitivity Test Information Matrix Functions for a Single Sample. The solid curve, $J_0(z)$, gives the information for the variance of $\hat{\mu}$; the dash-dot curve, $J_2(z)$, gives the variance of $\hat{\sigma}$, and the dash curve, $J_1(z)$, gives the covariance.

the mean and that the asymptotic variance of $\hat{\sigma}$ will be minimized by concentrating the testing at stimulus levels at $\mu \pm 1.6\sigma$ (for a normal distribution). Figure 1 also shows that it is impossible to simultaneously minimize the variances for both $\hat{\mu}$ and $\hat{\sigma}$.

For large sample sizes, the area of the standard confidence ellipsoid for the parameters is inversely proportional to the determinant of the information matrix. Since a *D*-optimal result will be obtained when the determinant of the information matrix is maximized, a *D*-optimal design gives the smallest confidence ellipsoid for the parameters. (See Silvey [1980], Wu [1985], and McLeish and Tosh (1990) for further discussion of optimality.) Since the off-diagonal terms of the matrix are typically small compared to the diagonal terms, a *D*-optimal test will also approximately minimize the product of the asymptotic variances of μ and σ. This condition is achieved (for a normal distribution) by testing near $\mu \pm 1.138\sigma$. (See also Banerjee 1980).

2. THE NEW TEST

Three sensitivity tests are most commonly used in the explosive-test community for evaluating explosives. In the probit (Bliss 1935) test, the experimenter chooses several stimulus levels and the number to be tested at each level before the test begins. The Bruceton test (Dixon and Mood 1948) requires the experimenter to supply an initial estimate of the mean, x_1, and a step size, d, close to the estimate of the standard deviation. The first specimen is tested at x_1. Specimen number $n + 1$ is tested at level $x_n - d$ if specimen n responded and at level $x_n + d$ if specimen n failed to respond. The Langlie test (Langlie 1965) requires the experimenter to specify a lower and an upper stress limit. The first test is conducted at a level midway between these limits. The remaining levels can be found by the prescription given by Langlie (1965, p. 12): "The general rule for obtaining the $(n + 1)$st stress level, having completed n trials, is to work backward in the test sequence, starting at the nth trial, until a previous trial (call it the pth trial) is found such that there are as many successes as failures in the pth through nth trials. The $(n + 1)$st stress level is then obtained by averaging the nth stress level with the pth stress level. If there exists no previous stress level satisfying the requirement stated above, then the $(n + 1)$st stress level is obtained by averaging the nth stress level with the lower or upper stress limits of the test interval according to whether the nth result was a failure or success."

Since these procedures test specimens at stress levels across a wide range of the distribution, they provide good estimates of μ and provide reasonable estimates of σ, assuming that the experimenter has started the test with accurate guesses of the parameters. If not much is known about the population parameters, however, all of these tests "waste" samples by testing far from the mean. Figure 1 shows that essentially no information is obtained by testing more than three standard deviations from the mean. (The results of many "nonparametric" tests [Chung 1954; Cochran and Davis 1965; McLeish and Tosh 1990; Robbins and Monro 1951; Wu 1985] designed to estimate a single stress level can be analyzed by maximum likelihood methods to estimate μ and σ. The parameters optimal for estimating both μ and σ are often different from the parameters chosen to optimize estimation of one specific quantile.)

This new test has three parts. The first part of the new test algorithm is designed to "close in" on the region of interest (to within a few standard deviations of the mean) as quickly as possible. The second part of the test is designed to determine unique estimates of the parameters efficiently. The third part of the test continuously refines the estimates once unique estimates have been established. This test has many of the same characteristics as a previously proposed test (Neyer 1989) but is conceptually simpler.

Figure 2 shows a flow chart of the procedure used to pick the next stress level. The experimenter uses his knowledge of the specimens to guess a lower and an upper bound for the mean ($\mu_{\min}$ and $\mu_{\max}$) and a guess of the standard deviation (σ_{guess}).

The first part of the test uses a modified binary search to get close to the mean. The first specimen is tested at level x_1, located midway between the $\mu_{\min}$ and $\mu_{\max}$. If the first specimen responds, then x_2 is the average of x_1 and $\mu_{\min}$. However, x_2 is set equal to $x_1 - 2\sigma_{\text{guess}}$ if that lies lower. A nonresponse of the first test is treated analogously. If the first two or more results are the same, the next level is chosen such that the range of stresses tested doubles with each test (i.e., $x_{n+1} - x_n = x_n - x_1$). Once at least one success and failure have been obtained, a binary search is performed until the difference between the lowest success and the highest failure is less than the estimate for sigma. This first part of the test was designed to yield both successes and failures quickly when the initial guesses were accurate and to expand rapidly when the estimates were in error. It is not possible to estimate the efficiency of this part of the test in all cases because the efficiency depends on the accuracy of the initial guesses. If the initial guesses are accurate and the range is smaller than $8 \times \sigma_{\text{guess}}$, however, the simulation reported later in this work demonstrated that this part of the test usually requires two samples.

The second part of the test is designed to provide unique estimates of the MLE's quickly. Unique es-

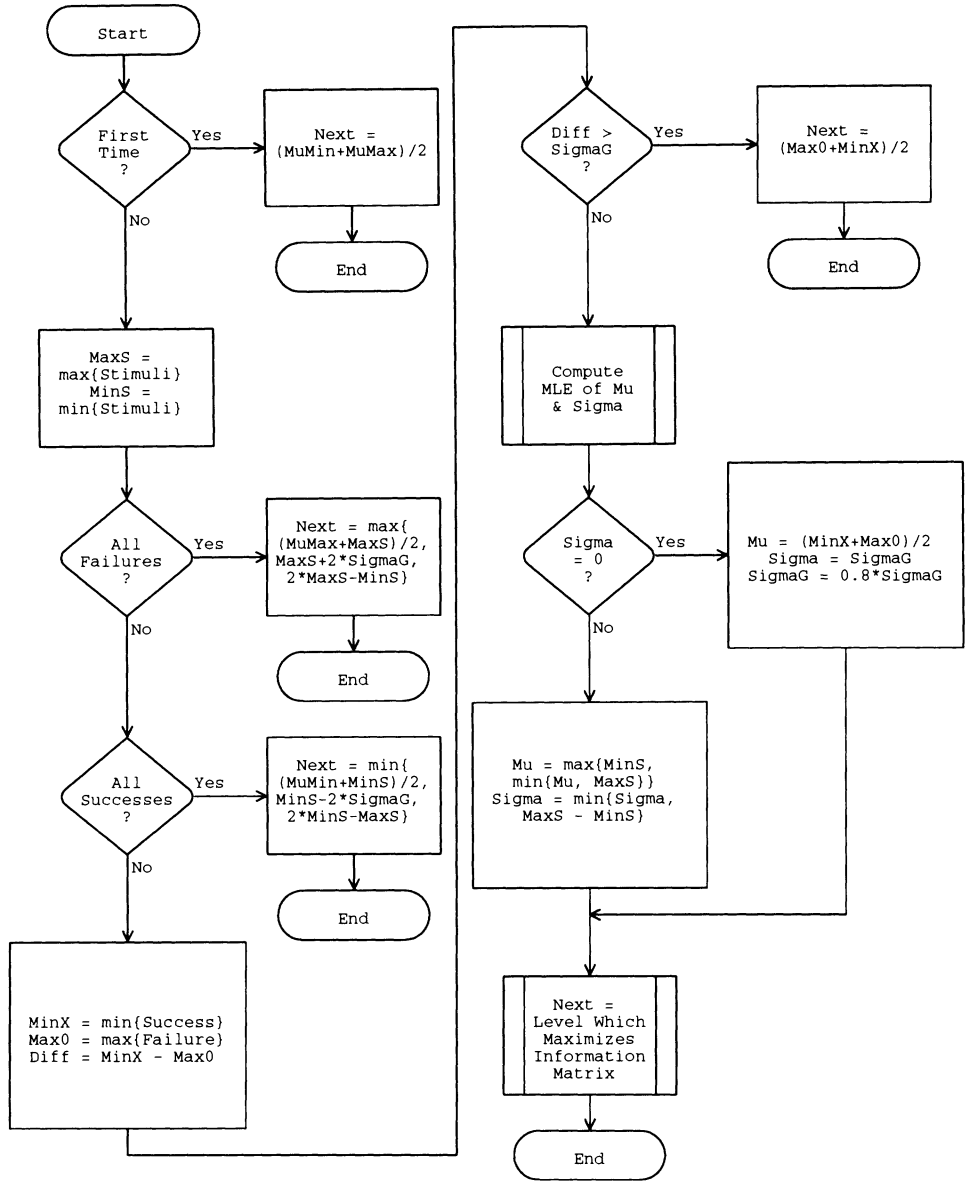

Figure 2. Flow Chart Showing the Algorithm Used to Determine the Next Test Level.

timates are achieved when the successes and failures overlap. The average of the lowest success and highest failure is used as an estimate of the mean, and σ_{guess} is used as an estimate of sigma. (These inaccurate estimates will only be used until the successes and failures overlap.) The next test level is chosen as that level that maximizes the determinant of the information matrix given these estimates. Thus this part of the test is similar to the initial part of the *c*-optimal design of McLeish and Tosh (1990). In the design proposed here, however, μ_{guess} is updated and σ_{guess} is decreased by multiplying by .8 for each specimen tested. Decreasing sigma results in faster overlap of the data. It also prevents the procedure from testing all specimens far from the mean when $\sigma \ll \sigma_{\text{guess}}$. The value of .8 used to multiply σ_{guess} was chosen from the results of several simulations. Smaller values improved the efficiency slightly when and only when $\sigma \ll \sigma_{\text{guess}}$. Larger (up to .85) values improved the efficiency when $\sigma \geq \sigma_{\text{guess}}$ but significantly decreased the efficiency when $\sigma \ll \sigma_{\text{guess}}$. Since this value is only used to quickly determine unique estimates of the MLE's, it has no effect on the marginal efficiency of the test.

The final part of the test is similar to the second, except that the MLE's are used as estimates of the parameters. Unfortunately, the MLE's will sometimes be "wild" estimates when computed from a limited number of tests. Thus the algorithm limits $\hat{\mu}$ to lie within the range of the stimulus levels tested previously and limits $\hat{\sigma}$ to be less than the difference between the highest and lowest levels tested. In such

a case, the next test is usually outside the test range; thus the limits are expanded (usually more than doubled) so that the limits will not constrain the true parameters. Since wild estimates generally occur in much less than 1% of cases, no attempt was made to find the optimal restriction. Limiting the estimates of the parameters has a similar effect as the truncated version of Wu's test (Wu 1985).

The algorithm was designed to be "fail-safe"; even if the mean is far outside the specified range, the first part of this algorithm will expand to contain and then converge to the region of interest. It will produce unique estimates for σ, even if the true σ is a factor of 10 or more smaller, assuming that the sample size is sufficient.

The following example should clarify the algorithm used in the new test. Suppose that a company manufactures an explosive and tests it with an explosive drop-weight test. Assume that a regular batch of explosives is normally distributed and has a mean drop height threshold of 1 meter (m) and a standard deviation of .1 m. When testing a regular batch of explosives, the experimenters perform a 20-sample test, with variables $\mu_{\min} = .6$ m, $\mu_{\max} = 1.4$ m, and $\sigma_{\text{guess}} = .1$ m. Now suppose that a batch of explosives was improperly mixed, raising its mean threshold to 5 m and its standard deviation to 1 m.

Table 1. An Example of the New Test When Sample Is Different Than Expected

No.	Drop height (m)	Result	Comment
1	1.00	Failure	Start with binary search.
2	1.20	Failure	
3	1.40	Failure	Regular explosive would have exploded!
4	1.80	Failure	Rapidly increase upper limit to get success quickly.
5	2.60	Failure	
6	4.20	Success	Both successes and failures! Begin binary search.
7	3.40	Failure	
8	3.80	Failure	
9	4.00	Failure	
10	4.10	Failure	
11	4.28	Failure	(No overlap. Use $\hat{\mu} = 4.15$, $\hat{\sigma} = \sigma_{\text{guess}} = .10$.)
12	4.52	Failure	$\hat{\mu} = 4.28$, $\hat{\sigma} = .19$. (Overlap. Clip MLE values.)
13	5.55	Success	$\hat{\mu} = 4.52$, $\hat{\sigma} = .79$.
14	5.24	Failure	$\hat{\mu} = 4.66$, $\hat{\sigma} = .50$. (Use true MLE values.)
15	6.37	Success	$\hat{\mu} = 5.22$, $\hat{\sigma} = .96$.
16	6.08	Failure	$\hat{\mu} = 5.10$, $\hat{\sigma} = .83$.
17	7.38	Success	$\hat{\mu} = 5.70$, $\hat{\sigma} = 1.39$.
18	7.09	Success	$\hat{\mu} = 5.58$, $\hat{\sigma} = 1.25$.
19	6.89	Success	$\hat{\mu} = 5.50$, $\hat{\sigma} = 1.16$.
20	6.74	Success	$\hat{\mu} = 5.44$, $\hat{\sigma} = 1.09$.

If the experimenters did not know about the improper mixture and conducted the test using the new test procedure, with the variables appropriate to a regular batch of explosives, the experiment would yield results similar to those shown in Table 1.

Analysis of these data yields MLE's $\hat{\mu} = 5.39$ m and $\hat{\sigma} = 1.04$ m. Thus, even though the defective batch of explosives was very different from a regular batch, the new sensitivity test quickly led to good estimates. Many of the standard sensitivity tests in common use would have wasted their specimens by testing far below the mean.

It would be possible to modify other tests to efficiently search for the region of interest. Langlie (1990, personal communication) suggested shifting the test range up (down) by $\frac{1}{3}$ of its size when a failure (success) occurs in the upper (lower) 70% of the test range. Instead of using this approach, many experimenters decrease the efficiency of the Langlie test by expanding the test range to ensure that all specimens will respond at the upper limit and fail at the lower limit. It is also possible to use larger step sizes in a Bruceton test until the region of interest is reached, although a very large step size would result in an inefficient test. The first parts of this procedure could also be used as a starting algorithm for many of the *c*-optimal tests.

3. COMPARISON OF TECHNIQUES

Several authors (Davis 1971; Edelman and Prairie 1966; Langlie 1965; Wu 1985) have compared various sensitivity tests to determine the most efficient method for estimating parameters of the distribution. A Monte Carlo approach similar to that used in several of the previous works was used to compare the new test described here to the probit, Bruceton, and Langlie tests commonly used in explosive research. No comparison is reported here with any *c*-optimal designs due to space limitations. Further simulation and previous work (Neyer 1989) show that the new test is more efficient at estimating σ than the *c*-optimal designs studied although less efficient at estimating μ.

There is an optimal condition for conducting each of the test designs. The preceding analysis shows that a probit test will be *D*-optimal if the stress levels are evenly distributed near the two points $\mu \pm 1.138\sigma$. The optimal condition for the Bruceton test is to pick the step size equal to the standard deviation (Dixon and Mood 1948). Langlie (1965) suggested limits of $l_{\min} = \mu - 4\sigma$ and $l_{\max} = \mu + 4\sigma$. Simulation has shown that the optimal condition for the new test is for $l_{\min} \geq \mu - 4\sigma$, $l_{\max} \leq \mu + 4\sigma$, and $\sigma_{\text{guess}} = \sigma$. Since the reason for performing any test is to determine the parameters, any realistic comparison of tests must include simulation under a variety of conditions.

For the simulations to be reported, a mean, μ_{guess}, and a standard deviation, σ_{guess}, were given as initial guesses for all of the sensitivity tests. The tests were optimized for this choice of parameters; if the true parameters agreed with the initial guesses, then the tests should be most efficient. The tests were performed with several different values of the true mean, μ, and standard deviation, σ. Monte Carlo simulations of 10,000 repetitions were performed for sample sizes from 6 to 50. Simulations are performed with $\mu \approx \mu_{guess}$ and $\sigma = (.1, .2, .5, 1.0, 2.0, 5.0, 10.0)\sigma_{guess}$. Additional simulations were performed with $\mu \approx \mu_{guess} + (1, 2, 5, 10)\sigma_{guess}$ and $\sigma = (.2, .5, 1.0)\sigma_{guess}$. Combinations with μ offset and large values of σ were not performed for most tests because the large σ masked the effect of μ offsets.

Although the experimenter rarely knows the exact value of the mean before the test, the efficiency of the Bruceton test is different when the first test level is exactly on the mean than when it is offset by $\sigma/2$ (Dixon and Mood 1948). The efficiency of the Langlie test is a complicated function of the position of the mean with respect to the test limits (H. J. Langlie, 1990, personal communication). The value of μ used for each test was offset by a different random number between $\pm .5\sigma$ to model experimenter uncertainty of the exact position of the mean. (Additional simulations keeping μ fixed showed that, although the efficiency of the new test was independent of the exact position of μ relative to the test parameters, the efficiency of the Bruceton and Langlie tests was dependent.)

To compare the efficiencies, MLE's of μ and σ were computed for each test. From these, the simulation mean squared errors (MSE's) of both $\hat{\mu}$ and $\hat{\sigma}$ were computed for each set of initial parameters. Because several of the MLE's were "wild," truncated MLE's were used for computing the MSE's. If $\hat{\sigma} > 5\sigma$, it was replaced with 5σ, and if $\hat{\mu} > \mu + 5\sigma$ or $\hat{\mu} < \mu - 5\sigma$, it was replaced with the closer of $\mu \pm 5\sigma$, where σ is the true standard deviation of the population. This truncation scheme is used only in the computation of the MSE's and not in the algorithm that picks the next stress level. (See McLeish and Tosh [1990] for another method of eliminating the wild MLE's.)

Figures 3–7 show σ^2/MSE as a function of the sample size for both μ and σ for all of the tests studied. Due to space limitations, only 5 of the 19 combinations of μ and σ are shown. The other combinations showed similar results. Shown also is σ^2 divided by the asymptotic variance, assuming that the tests were conducted at the D-optimal test points $\mu \pm 1.138\sigma$. (This function is a straight line approximated by $.392N$ for μ and $.507N$ for σ. The coefficients, derived by Banerjee [1980], can be read from Fig. 1.) The curves for the Bruceton, Langlie, and new tests all use the random offset value of the mean to simulate experimenter uncertainty about the exact value of the mean. Figure 3 also shows the MSE's for the probit test with no random offset. The probit test is not shown on the other graphs because the value of σ^2/MSE is close to 0. (Since there are several "wild" estimates when $\sigma \gg \sigma_{guess}$, the μ efficiency

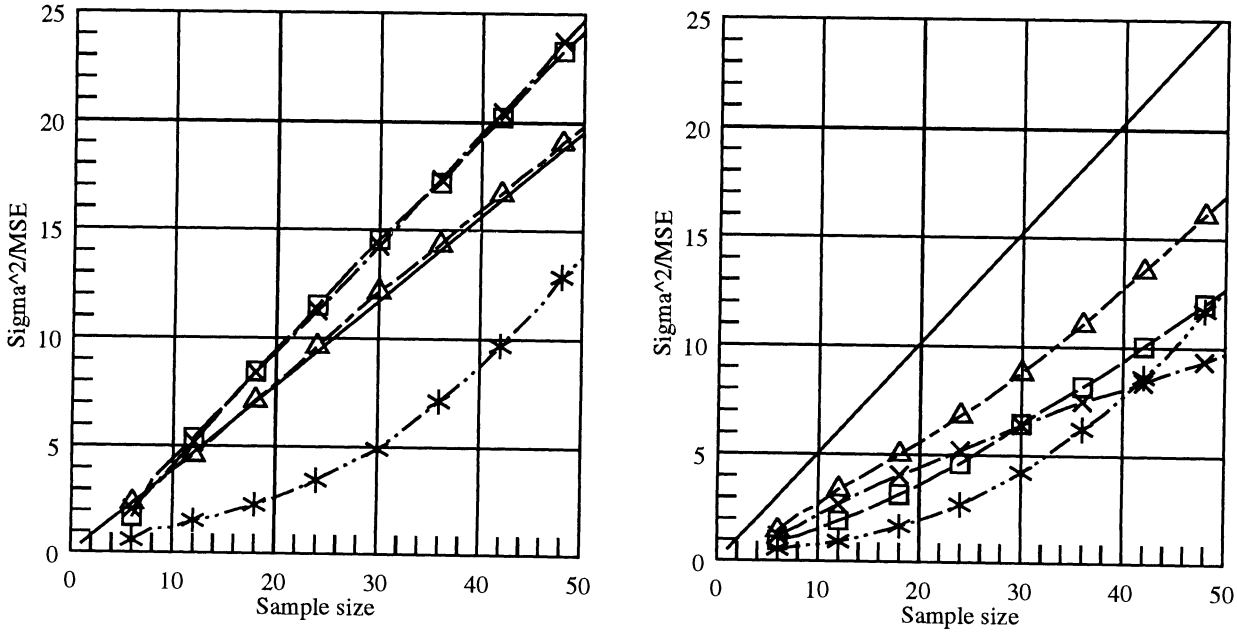

Figure 3. Efficiency in Determining μ (left) and σ (right) for the Various Test Designs When $\mu \approx \mu_{guess}$ and $\sigma = \sigma_{guess}$: ——, Ideal; –□–, Bruceton; –·X–·, Langlie; –·△–·, New; ––··, Probit.*

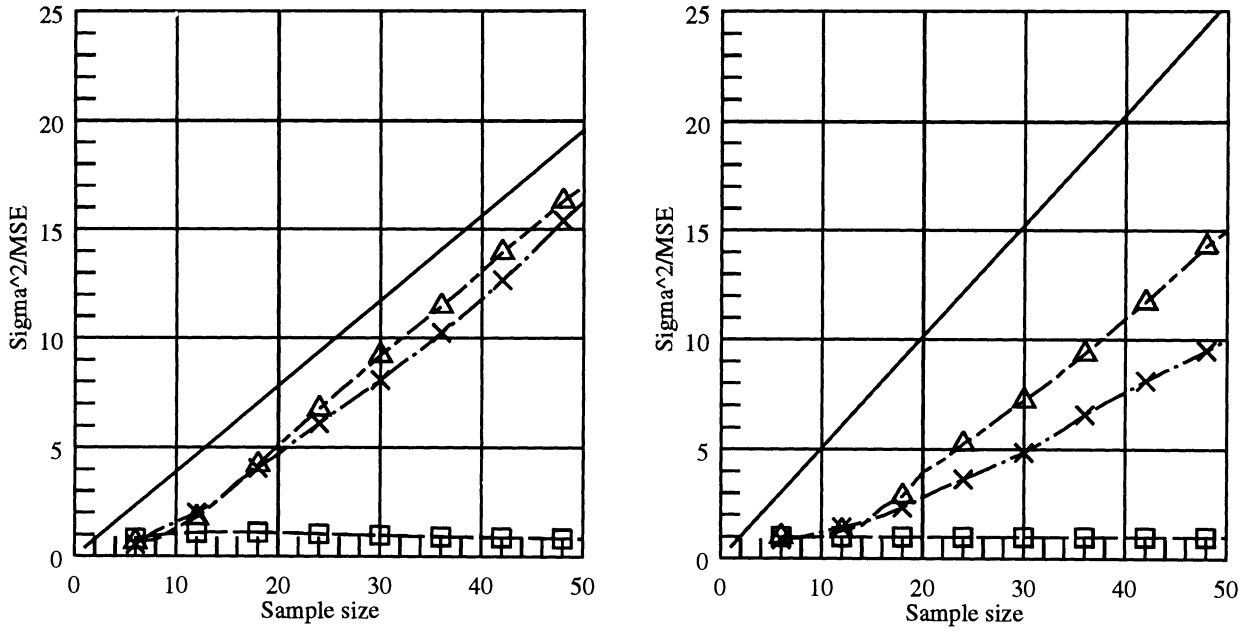

Figure 4. Efficiency in Determining μ (left) and σ (right) for the Various Test Designs When $\mu \approx \mu_{guess}$ and $\sigma = .2\sigma_{guess}$: ———, Ideal; –□–, Bruceton; –·X–·, Langlie; –·△–·, New.

curves are not smooth for the Bruceton and Langlie tests in Fig. 5.)

Several conclusions can be drawn from the simulation. The new test provides much better estimates of the standard deviation than the other tests under all the conditions tested. The efficiency in determining the mean is worse for the new test compared to the Bruceton and Langlie tests when the parameters are close to the initial guess but is comparable to or better than the other tests when the initial guesses of the parameters are far from the true values.

The efficiency of a test was not strongly dependent on a wrong guess for μ for any of the tests studied. Offsets of two σ or less had little effect. Larger μ offsets severely limited the efficiency of Langlie tests but only resulted in "wasting" the initial samples for the other tests. The efficiency was more strongly dependent on the estimate of σ. If the true σ was smaller

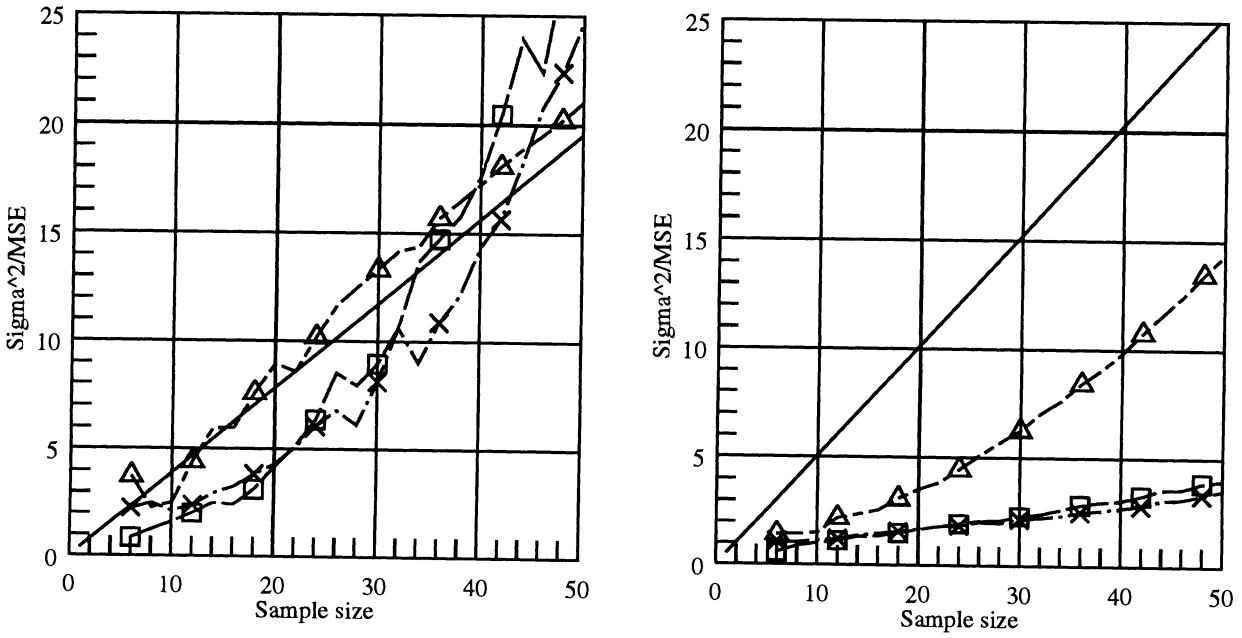

Figure 5. Efficiency in Determining μ (left) and σ (right) for the Various Test Designs When $\mu \approx \mu_{guess}$ and $\sigma = 5\sigma_{guess}$: ———, Ideal; –□–, Bruceton; –·X–·, Langlie; –·△–·, New.

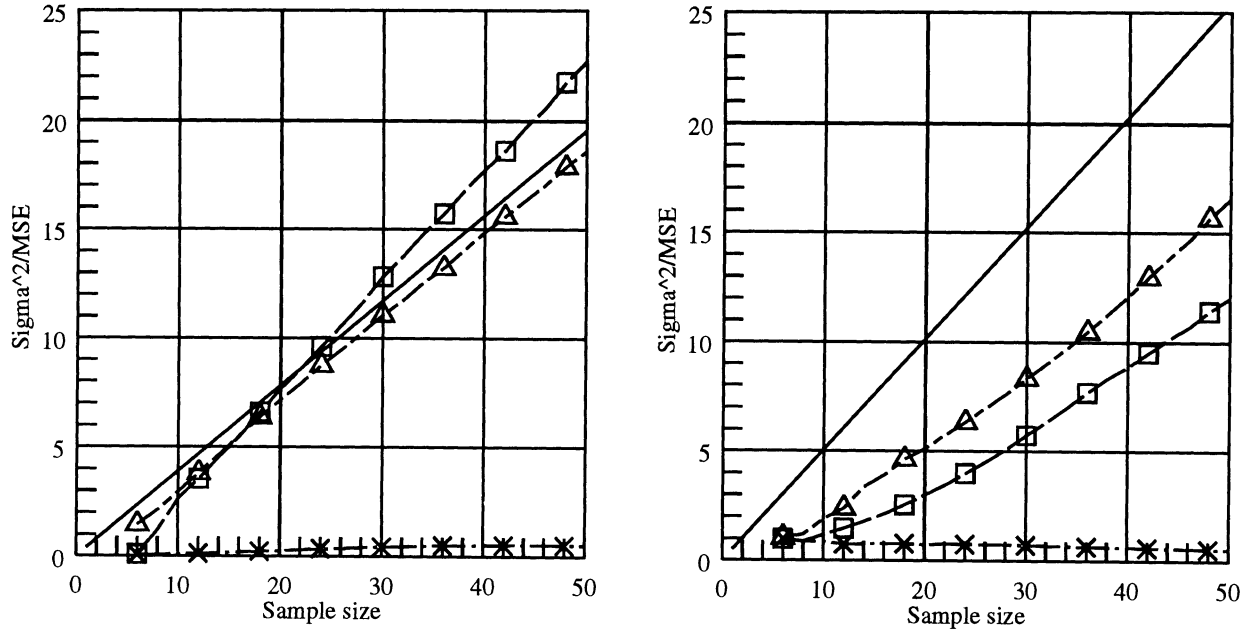

Figure 6. Efficiency in Determining μ (left) and σ (right) for the Various Test Designs When $\mu \approx \mu_{guess} + 5\sigma_{guess}$ and $\sigma = \sigma_{guess}$: ———, Ideal; –□–, Bruceton; –·X–·, Langlie; –·△–·, New.

than one-half of σ_{guess}, the Bruceton test was extremely inefficient for estimating both parameters. If the true σ was larger than guessed, the efficiency for estimating μ changes little for all tests studied. All tests but the new test, however, were less efficient when estimating σ when σ was larger than guessed. Since all graphs for the new test are approximately parallel to the asymptotic function, the marginal efficiency of the new test is close to 100%. The only effect of incorrect guesses for the parameters is inefficient initial testing.

The bias in estimating the parameters was also established with the simulation. There was no large bias in the MLE's for the mean for any design studied. All of the test designs produced biased estimates of σ, however. Figure 8 shows the relative bias of the standard deviation for the tests as a function of the sample size, N, under the conditions that the

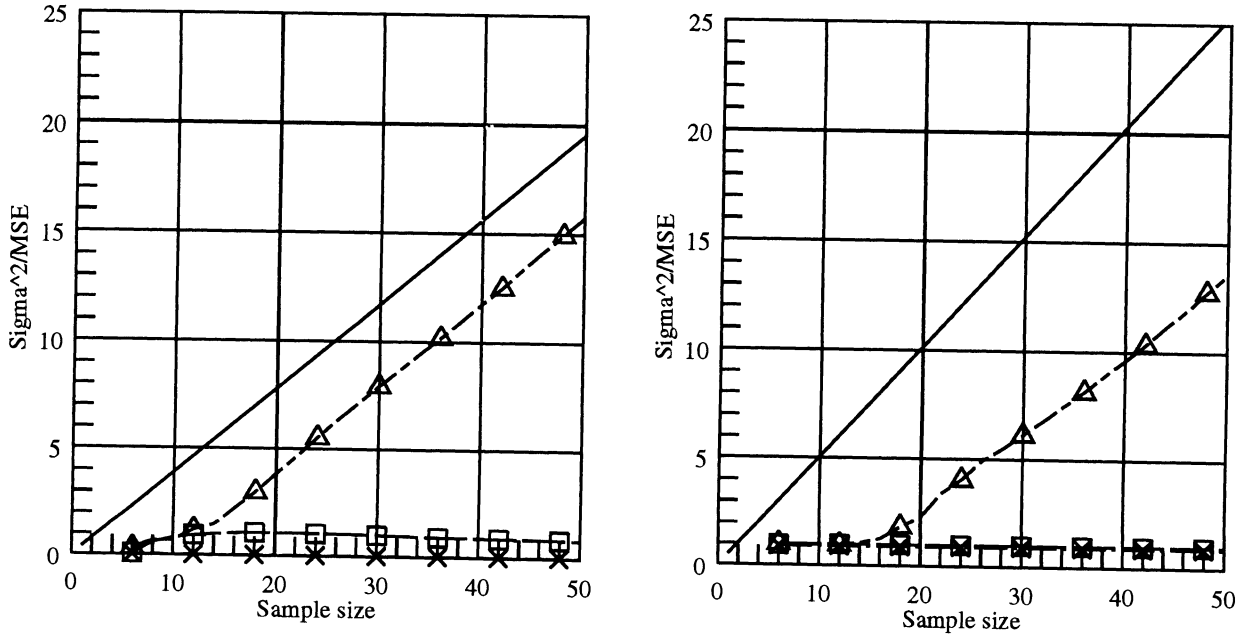

Figure 7. Efficiency in Determining μ (left) and σ (right) for the Various Test Designs When $\mu \approx \mu_{guess} + 5\sigma_{guess}$ and $\sigma = 5\sigma_{guess}$: ———, Ideal; –□–, Bruceton; –·X–·, Langlie; –·△–·, New.

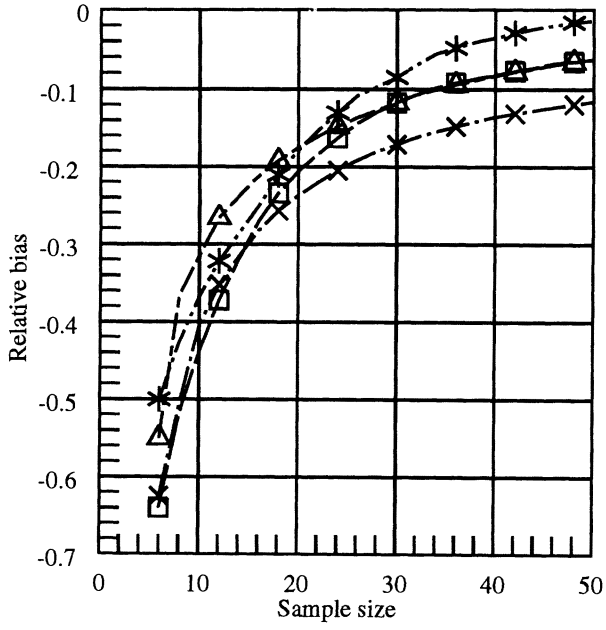

Figure 8. Relative Bias in the Maximum Likelihood Estimator for σ for the Various Test Designs When $\mu \approx \mu_{guess}$ and $\sigma = \sigma_{guess}$: –□–, Bruceton; –·X–·, Langlie; –·△–·, New; –·–, Probit.*

guessed parameters correspond to the true parameters. The magnitude of the bias decreased for all tests as the sample size increased. The relative bias for the new test can be approximated by the equation

$$\text{relative bias} \approx \frac{-3.5}{N}, \qquad (7)$$

determined by fitting a straight line to the reciprocal of the bias. Under nonideal conditions, the magnitude of the bias was much larger for the Bruceton and probit tests, slightly larger for the Langlie test, and essentially the same for the new test. Since the bias was small compared to the square root of the variance of the standard deviation for the new test, it contributed little to the MSE of σ.

As mentioned previously, each of the tests occasionally produced wild estimates for both μ and σ. The probit test was the most susceptible to these errors. Under ideal conditions with a sample size of 20 and 10,000 repetitions of the simulation, the probit test had 17 wild values and the Bruceton, Langlie, and new tests had none. If the guess for σ was too small by a factor of 5, the number of wild estimates were as follows: The Bruceton test, 149; the Langlie test, 102; and the new test, 2. For the probit test with an error in σ of a factor of 2, the corresponding number of wild values was 559. Thus, in addition to being more efficient at determining both μ and σ, the new test also has a much lower probability of producing wild estimates of these parameters.

The simulation showed that the new test was robust with respect to wrong initial guesses of the parameters. The simulation, however, was based on the assumption that the (transformed) stimulus levels were distributed normally, making it impossible to study robustness of the test to improper model specification. Young and Easterling (1994) studied a number of tests, including this *D*-optimality-based test, to determine their efficiency of estimating extreme quantiles. They found that tests (such as this one) designed to estimate parameters of the distribution gave more accurate quantile estimates than tests designed to estimate a single quantile but that they were more affected by model misspecification.

4. OTHER VERSIONS OF THE TEST

Several other versions of this test have been tried. One version of the test that has proved useful is a modification of the initial binary search. If the experimenter encounters three successes in a row, then he tests the next specimen at the highest failure minus half the distance between the highest failure and the lowest success. Three consecutive failures yield a similar modification. This choice for an initial search is slightly less efficient when the guesses are close to the true parameters but is much more efficient in the case when $\sigma \gg \sigma_{\text{guess}}$.

Simulations were also performed using Langlie and Robbins–Monro tests as the initial part of a *D*-optimality-based test design. These initial designs yielded slightly less efficient tests under ideal conditions and much less efficient designs when the initial guesses were different from the true parameters.

The binary search can be further modified if extra information is available to the experimenter. For example, the length of time a specimen survives before dying or how sick a specimen becomes before recovering can be used to give an indication of how close the stress was to the threshold in many biology experiments. If the experimenter has any knowledge of the relationship of the response to the distance from threshold, then this information could be used to generate a more efficient initial search. It could also be incorporated into the *D*-optimality-based part of the algorithm.

This new test is designed using the notion of *D*-optimality. Other optimizations would also be possible, however. Minimizing the asymptotic variance of a quantile L_p would result in a *c*-optimal design. This design would be similar to the test proposed by McLeish and Tosh (1990). Additional simulation, however, has shown it to be slightly more efficient when the guesses correspond to the true parameters and much more efficient when the guesses are different because their design kept μ and σ fixed until the successes and failures overlapped. In addition, the procedure could easily be modified using *E*-optimality or *G*-optimality criteria. Maximizing the

σ term in the information matrix would yield a design that would efficiently determine σ and extreme levels.

Earlier versions of this test (Neyer 1989) used a modified "memoryless" E-optimality-based design. (See McLeish and Tosh [1990] for a discussion on "memoryless min–var" design.) This design minimized the maximum variance of μ and σ. It gave similar results to the test described in this article.

5. PRACTICAL CONSIDERATIONS

Often before a test begins an experimenter needs to know the sample size needed to obtain a required precision. Since the precision scales with the standard deviation, it is not possible to specify the sample size needed to achieve a given precision such as 1 m. It is possible, however, to estimate the size necessary to determine a parameter to within a given fraction of σ. If all of the tests were conducted at $\mu \pm 1.138\sigma$, then the asymptotic variances for μ and σ would be given by $\text{var}(\mu) = \sigma^2/(.392N)$ and $\text{var}(\sigma) = \sigma^2/(.507N)$.

Inspection of Figure 3 shows that the variance of μ is close to the asymptotic variance, but the variance of σ is closer to the curve $\text{var}(\sigma)_{\text{real}} \leq \sigma^2/(.507(N - 15))$ for larger values of N. Further simulations with sample sizes up to 500 have shown that the same bounds apply. Thus the experimenter should ensure that the sample size is large enough to make sure that the required variance is as small as necessary.

The other figures show that the experimenter should add two or three extra specimens for each factor-of-two uncertainty in the knowledge of σ and one extra specimen for each factor-of-two increase in the range of possible values of μ beyond 8 σ. The number is only true on the average. Thus the experimenter should include extra samples if a certain precision is necessary.

6. SUMMARY

A new class of sensitivity tests has been proposed to determine the parameters of the underlying distribution. Theoretical analysis suggests and simulation has shown that the new test is able to efficiently determine both parameters of the distribution under all circumstances tested. If these tests are used, the experimenter can achieve greater precision and accuracy with the same sample size.

ACKNOWLEDGMENTS

I thank Kathleen Diegert of Sandia National Laboratories, Albuquerque, New Mexico, and Linda Young of the Biometry Department of the University of Nebraska, Lincoln, Nebraska, for their many suggestions and patient answers to my many questions. I also thank the reviewers for their helpful suggestions. EG&G Mound Applied Technologies is operated for the U.S. Department of Energy under Contract DE-AC04-88DP43495.

[*Received June 1989. Revised April 1993.*]

REFERENCES

Banerjee, K. S. (1980), "On the Efficiency of Sensitivity Experiments Analyzed by the Maximum Likelihood Estimation Procedure Under the Cumulative Normal Response," Technical Report ARBRL-TR-02269, U.S. Army Armament Research and Development Command, Aberdeen Proving Ground, MD.

Bliss, C. I. (1935), "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology*, 22, 134–167.

Chung, K. L. (1954), "On a Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 25, 463–483.

Cochran, W. G., and Davis, M. (1965), "The Robbins–Monro Method for Estimating the Median Lethal Dose," *Journal of the Royal Statistical Society*, Ser. B, 27, 28–44.

Cornfield, J., and Mantel, N. (1950), "Some New Aspects of the Application of Maximum Likelihood to the Calculation of the Dosage Response Curve," *Journal of the American Statistical Association*, 45, 181–210.

Davis, M. (1971), "Comparison of Sequential Bioassays in Small Samples," *Journal of the Royal Statistical Society*, Ser. B, 33, 78–87.

Dixon, J. W., and Mood, A. M. (1948), "A Method for Obtaining and Analyzing Sensitivity Data," *Journal of the American Statistical Association*, 43, 109–126.

Edelman, D. A., and Prairie, R. R. (1966), "A Monte Carlo Evaluation of the Bruceton, Probit, and One-Shot Methods of Sensitivity Testing," Technical Report SC-RR-66-59, Sandia Corporation, Albuquerque, New Mexico.

Golub, A., and Grubbs, F. E. (1956), "Analysis of Sensitivity Experiments When the Levels of Stimulus Cannot be Controlled," *Journal of the American Statistical Association*, 51, 257–265; Corrigenda (1956) 51, 650–651.

Kendall, M. G., and Stuart, A. (1967), *The Advanced Theory of Statistics* (Vol. 2, 2nd ed.), New York: Hafner.

Langlie, H. J. (1965), "A Reliability Test Method For 'One-Shot' Items," Technical Report U-1792, Aeronutronic Division of Ford Motor Company, Newport Beach, California.

McLeish, D. L., and Tosh, D. (1990), "Sequential Designs in Bioassay," *Biometrics*, 46, 103–116.

Neyer, B. T. (1989), "More Efficient Sensitivity Testing," Technical Report MLM-3609, EG&G Mound Applied Technologies, Miamisburg, Ohio.

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407.

Silvapulle, M. J. (1981), "On the Existence of Maximum Likelihood Estimators for the Binomial Response Models," *Journal of the Royal Statistical Society*, Ser. B, 43, 310–313.

Silvey, S. D. (1980), *Optimal Design*, London: Chapman and Hall.

Wu, C. F. J. (1985), "Efficient Sequential Designs With Binary Data," *Journal of the American Statistical Association*, 80, 974–984.

——— (1988), *Optimal Design for Percentile Estimation of a Quantal Response Curve*, New York: North-Holland.

Young, L. J., and Easterling, R. G. (1994), "Estimation of Extreme Quantiles Based on Sensitivity Tests: A Comparative Study," *Technometrics*, 36, 48–60.