A simulation-based investigation of the staircase method for fatigue strength testing

R. Pollak *, A. Palazotto, T. Nicholas

Department of Aeronautics & Astronautics, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH 45433, United States

Received 8 August 2005; received in revised form 1 November 2005

Abstract

A critical evaluation of the statistics of the fatigue strength distribution as determined by the staircase (or up-and-down) method is presented. The effects of test parameters (namely, step size and sample size) were analyzed using numerical simulation to determine the accuracy of fatigue strength standard deviation calculations using traditional staircase statistics, resulting in a quantification of standard deviation bias as a function of step size and sample size. A non-linear correction was formulated to mitigate this standard deviation bias inherent in small-sample tests. In addition, the simulation was used to investigate the effectiveness of a bootstrapping algorithm on standard deviation estimates. The bootstrapping algorithm was found to significantly reduce the potential of large standard deviation errors in small-sample tests. Together, the use of the non-linear correction factor and the bootstrapping algorithm may allow an improved method to estimate the statistics of a material’s fatigue strength distribution using a small-sample staircase test strategy.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Staircase testing; Fatigue strength testing; Bootstrapping; Numerical simulation; Fatigue limit

1. Introduction

Engineering structures are designed to provide reliable function for a given design life. In the case of components subjected to very high cycle fatigue, one of the key design considerations is the statistical distribution of the fatigue strength of the material at the number of cycles the component is designed to withstand. This distribution can then be used to specify safe operating loads with minimal risk of fatigue failure. Several different testing strategies have emerged over the years to evaluate fatigue strength.

One relatively common approach to fatigue strength testing is the staircase method first statistically analyzed by Dixon and Mood (1948) for use in explosives testing, and later popularized by Little (1975a) for application to fatigue testing. The staircase method utilizes a simple protocol in which a specimen is tested at a given starting stress for a specified number of cycles or until failure, whichever comes first. If the specimen survives, the stress level is increased for the next specimen; likewise the stress is decreased if the specimen fails. This protocol is
continued for a batch of specimens with Dixon and Mood’s equations applied to the results in order to estimate the mean fatigue strength and its standard deviation at the specified number of cycles. The method is remarkably accurate and efficient in terms of quantifying the mean fatigue strength. Unfortunately, it is difficult in practice to provide accurate estimates of the standard deviation of the fatigue strength using this method for small-sample test programs typical of high-cycle or ultra high-cycle fatigue testing. Several recent investigations have been conducted to assess and improve the reliability of the standard deviation estimate using the staircase method, to include the work of Braam and van der Zwaag (1998), Svensson and de Maré (1999), Lin et al. (2001) and Rabb (2003). Of these investigations, the correction factor proposed by Svensson et al. is of special interest to this study as a starting point for reducing small-sample bias inherent in standard deviation estimation using the Dixon and Mood method.

The objective of this research was to both characterize the standard deviation bias inherent in staircase tests, and propose means to improve standard deviation estimates to allow use of the method for small-sample ultra high-cycle fatigue tests. A numerical simulation was designed which replicates a staircase test for a given test design and underlying true fatigue strength distribution. There were two primary considerations when evaluating the staircase methodology. The first is assessing how accurate the test is on average. In this sense, the goal is obviously to have a test method which estimates the central tendency of fatigue strength distribution parameters as closely to true values as possible – i.e., a test with little bias in results. The second consideration involves the scatter in results. Because a designer or researcher must make conclusions based on a limited set of data, it is important that a test should not be subject to a wide dispersion in results due to statistical scatter alone. In other words, one would prefer a test method with results as tightly grouped as possible around the true parameter values.

These two considerations led to two different techniques to improve the fatigue strength analysis using a small-sample staircase test strategy. The first technique involves the use of a non-linear correction term to the Dixon and Mood standard deviation estimate in order to mitigate the bias. The second technique is the incorporation of a simple bootstrapping algorithm to help mitigate scatter in results. Through a combination of these techniques, a modified staircase strategy emerges which is generally more accurate and precise than traditional Dixon and Mood analysis alone.

2. The staircase method

The staircase method is a quantal response test. As opposed to conventional stress-life (S–N) analysis, quantal response analysis does not consider the actual number of cycles to failure, but rather tests are handled in a “pass/fail” manner. Thus, quantal response tests are designed to handle runout test data (i.e., tests in which the specimen did not fail). Although used in fatigue strength testing and other applications, quantal response methods are historically associated with biological assay. Biological assay is a set of techniques used in comparisons of alternative but similar biological stimuli – basically, the measurement of the potency of any stimulus by observing the reaction that it produces in a living organism (Finney, 1971). The objectives of biological assay and fatigue strength testing are quite similar. In both cases, one wishes to determine the appropriate stimulus level (“stress” or “dose”) for which an acceptable proportion of specimens survive.

The staircase test was first proposed by Dixon and Mood (1948) for application to explosives testing, in which the stimulus was the drop height, and the response was either a detonation or failure to fuze. In a staircase test, specimens are tested sequentially, with the first specimen tested at an initial stress level (S₀), typically the best guess for median fatigue strength estimated from either experience or preliminary S–N data. The stress level for the next specimen is increased or decreased by a given interval depending on whether the first specimen survives or fails. This process is continued until all the specimens allocated for the experiment have been used. Typically, the step size (s) between adjacent stress levels is held constant (approximately equal to the standard deviation of fatigue strength), in which case the statistics of Dixon and Mood may be applied directly to estimate mean and standard deviation of the fatigue strength at a given number of cycles. Even though the true standard deviation in fatigue strength is one of the unknowns, it is not too important if the interval is actually incorrect with respect to the true standard deviation by as much as 50% for estimates of mean fatigue strength (Dixon, 1965). However, arbitrary spacing between
stress levels may be used when accompanied by a
probit-type analysis (Little, 1975b). In fact, tests
conducted with non-uniform spacing may be more
statistically efficient than uniform spacing; however,
the analysis becomes much more tedious and the
equations and tables derived for uniformly spaced
tests are no longer useful (Little, 1972). Tests with
non-uniform spacing were not considered. Fig. 1
illustrates the staircase method for a constant step
approach.

The primary condition of the Dixon and Mood
analysis is that the variate under consideration must
be normally distributed. In the case of fatigue test-
ing, this would imply that the fatigue strength
distribution must be normally distributed. This
condition would seem to be very restrictive and rule out
staircase testing for many materials; however, a
transformation of the stress values may be applied.
In addition to logarithmic transformations, power
transformations are often applied for such a pur-
pose. Squared and cubic transformations (as well
as other powers greater than 1) are used to reduce
negative skewness, while logarithmic and powers
less than 1 (such as a square-root transformation)
as well as negative reciprocals (−1/x) are often used
to reduce positive skewness (Neter et al., 1996).
Thus, the normality condition is not as restrictive
as it first appears. However, it is likely to often be
the case where there is not enough data to actually
estimate the distribution shape prior to conducting
a staircase test. Thus, one must either make a guess
as to the shape based on data from a similar mate-
rial and apply an appropriate transformation if
non-normal, or simply accept the initial assumption
of normality and conduct the test without a stress
transformation and then use the data from the
experiment itself to estimate the distribution.

Dixon and Mood’s second condition is that the
sample size must be large, on the order of 40–50
specimens or more. This condition ensures that
large sample theory, on which the analysis is based,
can be applied. However, additional research with
respect to sample size has shown that this condition
is actually unnecessary when testing for fatigue
strength mean. Brownlee et al. (1953) note that the
distribution mean using Dixon and Mood’s analysis
is reliable even in samples as small as 5–10.

The final condition of the Dixon–Mood ap-
proach is that the standard deviation of the normal
distribution must be roughly estimated prior to test-
ing. This condition is necessary because the equa-
tions on which the standard deviation estimate is
based assume that the step size is on the order of
0.5σ to 2.0σ where σ is the true standard deviation
of the fatigue strength. Thus, some knowledge of
the standard deviation is required in order to specify
the step size prior to testing. This requirement is not
as severe as it may appear because usually some
data points are available from other testing of the
material of interest or a similar material in order
to provide a rough initial estimate of standard
deviation.

With these assumptions, Dixon and Mood used
maximum-likelihood estimation techniques to ana-
lytically solve the problem of determining the mean
and standard deviation of the variate of interest
based on staircase test data. The results of their
analysis are summarized by the equations below
for fatigue strength mean (μ) and standard devia-
tion (σ), which may be considered the “traditional”

---

![Fig. 1. Notional staircase test data.](image-url)
Define $A = \sum_{i=0}^{\text{max}} m_i$, $B = \sum_{i=0}^{\text{max}} im_i$, $C = \sum_{i=0}^{\text{max}} i^2m_i$.

$$\mu = S_0 + s \left( \frac{B}{A} \pm 0.5 \right)$$

$$\sigma = 1.62 \cdot s \cdot \left( \frac{A \cdot C - B^2}{A^2} + 0.029 \right)$$

if $\frac{A \cdot C - B^2}{A^2} \geq 0.3$

or, $\sigma = 0.53 \cdot s$ if $\frac{A \cdot C - B^2}{A^2} < 0.3$

In these equations, the parameter $i$ is an integer denoting the stress level, with $i_{\text{max}}$ corresponding to the highest stress level in the staircase. If the majority of specimens failed, then the lowest stress level at which a survival occurs (i.e., runout) responds to the $i = 0$ level and $m_i$ corresponds to the number of specimens which survived each stress level. The next highest stress level would be the $i = 1$ level, and the stress level one above that would be $i = 2$, etc. If the majority of specimens survived the given number of cycles, then the lowest stress level at which a failure was observed is denoted as the $i = 0$ level and $m_i$ corresponds to the number of specimens which failed at each stress level. The stress value corresponding to the $i = 0$ stress level; note that this is not necessarily the same as the initial starting stress, $S_{\text{init}}$. As already defined, $s$ is the step size. The plus sign in the equation for $\mu$ is used when failures are the majority event, while the minus sign is used if survivals are the majority event.

3. The standard deviation problem

Standard deviation estimation has been problematic using staircase testing because of the very nature of the testing itself. By concentrating the majority of the data points near the mean, it is more difficult to get an accurate measure of dispersion. With few data points in the “tails” of the distribution (i.e., the areas of decreasing probability), standard deviation estimates become based on data near the center of the distribution. Therefore, slight differences in the central data can lead to very different measures of standard deviation as there is little data in the tails to help scope the dispersion. When there are many data points, even if centrally located, the dispersion can be more accurately measured since the underlying distribution is fixed. For this reason, one can overcome the difficulties of standard deviation estimation using the staircase test if a very large number of specimens are available for testing. However, using a very large number of specimens would defeat the purpose of doing staircase testing in the first place, which is to realize gains in efficiency versus balanced test strategies such as the probit method.

After many years of apparently little work being published on the staircase test, there has recently been considerable interest in this area, specifically with respect to standard deviation estimation, by a number of different researchers. One can reasonably speculate that these efforts are in part due to the following factors: (1) increasing interest in the probabilistic aspects of fatigue, primarily high-cycle fatigue, (2) dramatically improved processing power of personal computers to handle more complex digital simulations, and (3) greater emphasis on efficient test strategies to provide reliable estimates due to the high cost of materials and components and longer test times in the ultra high-cycle regime.

In 1999, Svensson and Maré published an analysis of the random features of the fatigue limit. In this paper, the results of simulation work from small-sample tests ($\leq 30$ specimens) were summarized. This simulation work suggested that the standard deviation estimates using Dixon–Mood analysis are biased. In a follow-up paper summarizing the research (Svensson et al., 2000), a linear correction factor was proposed (here called the Svensson–Lorén correction) and found to be “an improvement in all maximum-likelihood evaluation procedures, including the staircase method”. Their equation is shown below, where $\sigma_{\text{SL}}$ represents the Svensson–Lorén corrected standard deviation estimate, $\sigma_{\text{DM}}$ is the standard deviation estimate based on Dixon and Mood analysis, and $N$ is the total number of specimens:

$$\sigma_{\text{SL}} = \sigma_{\text{DM}} \frac{N}{N - 3}$$

The correction is strictly a function of the sample size and has the effect of increasing the standard deviation estimate found from Dixon–Mood analysis, with the increase greater for smaller sample sizes. Since the standard deviation estimate is merely multiplied by a constant greater than 1, this correction will yield slightly more scatter in results compared to Dixon–Mood.
4. The staircase simulation

A computer simulation was designed to analyze the effects of staircase parameter settings on standard deviation estimates using the Dixon and Mood analysis method. The simulation allows the user to specify the true fatigue strength distribution (modeled as normal in all cases) by specifying the mean and standard deviation of the fatigue strength at the number of cycles of interest. The staircase test parameters (starting stress, step size, and number of specimens) are specified for each simulation run. For each specimen, the simulation calculates a random fatigue strength based on the specified underlying distribution and compares this value to the current stress level (starting with the initial stress, $S_{\text{init}}$) to determine if the specimen failed or survived. The stress level is increased or decreased for the next specimen according to the staircase protocol. This procedure is repeated until the total number of specimens is reached. Fatigue strength parameters are calculated according to Eqs. (1)–(4). This procedure is then repeated a large number of times (default 1000 runs) in order to provide a distribution of calculated standard deviations from which to draw conclusions. An initial comparison of results was accomplished for simulations made with 1000 runs versus 10,000 runs, showing that 1000-run simulations had adequate statistical confidence.

5. Effect of starting stress on standard deviation estimates

The first question of interest is what effect the initial starting stress has on standard deviation estimates. There are already adequate means of correcting for initial starting stress when calculating mean fatigue strengths (Brownlee et al., 1953; Dixon, 1965; Little, 1972). Furthermore, Rabb’s simulation work (2003), although using a different maximum-likelihood method from the Dixon and Mood method, showed that initial starting stress has little influence on standard deviation estimates. However, these simulations only used starting stresses at the mean or one-half step higher than the mean.

To address the effect of starting stress more rigorously, simulations were accomplished using an underlying Normal distribution with mean ($\mu$) equal to 400 and standard deviation ($\sigma$) equal to 5 (units are arbitrary). The step size was set equal to the true standard deviation. Three starting stresses were used: $\mu$, $\mu + 2\sigma$, and $\mu - 2\sigma$. The mean standard deviation estimates are plotted as a function of sample size in Fig. 2.

The data show that as sample size increases, starting stress becomes a non-factor in determining standard deviation, as one would expect. Even for small samples, the effect of starting stress is rather small. However, use of offset starting stresses does...
have some beneficial effect for small-sample sizes by alleviating some of the standard deviation bias. This result is due to the fact that when one starts above or below the mean, it is statistically more likely that a failure will be observed below the mean (if starting below) or a survival will be encountered above the mean (if starting above), simply because one is conducting more testing above or below the mean, on average. In a small-sample test, just one of these outcomes can lead to a higher estimate of standard deviation, and therefore on average, there is a slight reduction of the standard deviation underestimating bias for small-sample tests. This effect is not overwhelming enough to eliminate the bias altogether, but does provide some means of softening it. For the data in Fig. 2, the mean standard deviation error for \( N = 10 \) samples is 26.3% when starting at the mean, but 18.9% when starting two steps below the mean. The traditional approach of using \( S_{\text{init}} = \mu \) for staircase tests in order to maximize accuracy of the estimate for mean fatigue strength may be modified in order to improve standard deviation estimation, especially since adequate means of handling offset starting stresses exist for mean fatigue strength estimation.

6. Effect of step size on standard deviation estimates

It has been commonly understood, stemming from Dixon and Mood’s original work, that step sizes should be on the order of 0.5\( \sigma \) to 2\( \sigma \) when using the staircase method. Use of steps greater than 2\( \sigma \) results in staircases which tend to bounce back and forth across the mean but do not include enough stress level data with non-zero or non-unity probabilities of survival. When this situation occurs, Eq. (4) becomes applicable, and then the standard deviation estimate becomes a function of step size alone (0.5\( \sigma \)). Use of steps much smaller than 0.5\( \sigma \) risk requiring too many initial data points in order to walk up to or down to the mean in case the starting stress is significantly offset. In addition, use of very small steps results in significantly underestimated standard deviations. There has been some prior work in investigating the choice of step size, as Rabb (2003) looked at step sizes ranging from 0.85\( \sigma \) to 1.15\( \sigma \), while Braam and van der Zwaag (1998) investigated steps from 0.1\( \sigma \) to 1\( \sigma \). Neither work specifically recommended a step size, but left the impression that a 1\( \sigma \) step size is generally adequate. Both works, along with that of Svensson et al. (2000) and Svensson and de Mare ´ (1999), confirmed that standard deviation bias exists for small-sample tests.

The objective of this section was to quantify this standard deviation bias over the step size range of 0.1\( \sigma \)–2\( \sigma \). Simulations were run for both Normal(400,5) and Normal(400,15) underlying fatigue strength distributions. For each simulation, the starting stress was set at the true mean (i.e., \( S_{\text{init}} = \mu = 400 \)). Sample sizes ranged from 8 to 1000 specimens. When the standard deviation estimates are divided by the true standard deviation (i.e., normalized), the same outcome results irrespective of \( \sigma \). The normalized mean standard deviations for both the Normal(400,5) and Normal(400,15) cases were

![Fig. 3. Standard deviation bias for \( S_{\text{init}} = \mu \) using the Dixon and Mood method.](image-url)
averaged to create Fig. 3. This figure therefore gives the expected value of standard deviation \( \bar{\sigma}_{DM} \) using the Dixon and Mood method (for starting stresses at the true mean) for any combination of sample size \((\geq 8)\) and step size \((0.1\sigma–2\sigma)\). This figure shows that the Dixon and Mood method underestimates the standard deviation as either sample size or step size is reduced. In the limit (represented by 1000 specimens), the Dixon and Mood method is unbiased regardless of step size (in the 0.1\( \sigma \)–1.7\( \sigma \) region). Also, the method is generally unbiased for step sizes in the 1.6\( \sigma \)–1.75\( \sigma \) region. Beyond 1.5\( \sigma \), Eq. (4) begins to dominate as step sizes get large and the curves converge to the 0.53\( \sigma \) line.

Next, the scatter in standard deviation was addressed. Like \( \bar{\sigma}_{DM} \), the standard deviation of the standard deviation estimates \( \sigma_{DM} \) is independent of \( \sigma \) when normalized by \( \sigma \). Fig. 4 shows the average values of the normalized \( \sigma_{DM} \) estimates. Based on this data, step sizes in the unbiased region of 1.6\( \sigma \)–1.75\( \sigma \) produce less scatter in results than more traditional steps in the 0.5\( \sigma \)–1.5\( \sigma \) range. This effect is due in part to the larger steps producing more results requiring analysis by Eq. (4), which is constant for a given step size. In general, it appears that use of steps in the 1.6\( \sigma \)–1.75\( \sigma \) range result in both less bias and less scatter than smaller step sizes.

7. Standard deviation bias correction

Development of an approach to reduce standard deviation bias started with the Svensson–Lorén correction (Eq. (5)). The correction was applied to the normalized estimates shown in Fig. 3. This resulted in the normalized means of standard deviation using the Svensson–Lorén correction \( \bar{\sigma}_{SL} \) shown in Fig. 5. These results show a shift in the unbiased region from the 1.6\( \sigma \)–1.75\( \sigma \) range to the 0.85\( \sigma \)–1.0\( \sigma \) range. This shift allows relatively unbiased standard deviation estimates near 1\( \sigma \). A modified correction was developed which attempted to allow a greater range of unbiased estimation than the Svensson–Lorén correction. This proposed correction included the ratio of the standard deviation \( \sigma_{DM} \) to step size \( s \). The Svensson–Lorén sample size factor was included in order to retain the advantage of centering the unbiased region around the 1\( \sigma \) step size. The form of the proposed standard deviation estimate \( \sigma_{PC} \) is shown in Eq. (6), where \( A \), \( B \), and \( m \) are constants dependent on the number of samples (see Table 1).

\[
\sigma_{PC} = A \sigma_{DM} \left( \frac{N}{N-3} \right) \left( B \frac{\sigma_{DM}}{s} \right)^m
\]

The proposed correction was applied to the normalized Dixon and Mood mean estimates for standard deviation, as shown in Fig. 6. A comparison of normalized values of Dixon and Mood, Svensson–Lorén, and the proposed standard deviation estimates is shown in Fig. 7 for \( N = 20 \) specimens. These figures suggest that the proposed correction reduces bias over a wider range of step sizes. Note that these results are not dependent on the parameters of the underlying distribution (i.e., its mean and standard deviation), so long as it is normal. Unfortunately, the effectiveness of this correction is not
due to the effects of statistical scatter. It must be understood that these results are valid when applied to the average outcome. Staircase tests which produce results which are too high or too low may have errors magnified using this correction. The bootstrapping algorithm discussed next alleviates this condition to some degree.

8. Use of bootstrapping to reduce scatter

The bootstrap is a data-based simulation which utilizes multiple random draws from real test data to improve statistical inferences about the underlying population (Efron and Tibshirani, 1993). For the staircase data, the bootstrap concept was quite as significant as these figures would indicate, due to the effects of statistical scatter. It must be

Table 1

<table>
<thead>
<tr>
<th>Sample size (N)</th>
<th>A</th>
<th>B</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1.30</td>
<td>1.2</td>
<td>1.72</td>
</tr>
<tr>
<td>10</td>
<td>1.08</td>
<td>1.2</td>
<td>1.10</td>
</tr>
<tr>
<td>12</td>
<td>1.04</td>
<td>1.2</td>
<td>0.78</td>
</tr>
<tr>
<td>15</td>
<td>0.97</td>
<td>1.2</td>
<td>0.55</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>1.2</td>
<td>0.45</td>
</tr>
<tr>
<td>30</td>
<td>1.00</td>
<td>1.2</td>
<td>0.22</td>
</tr>
<tr>
<td>50</td>
<td>1.00</td>
<td>1.2</td>
<td>0.15</td>
</tr>
</tbody>
</table>
applied based on the associated probabilities of failure for each stress level, computed using the number of survivals and failures at each stress level. Using this data, simulated staircase tests can be generated numerically using the same starting stress, step size, and number of specimens. For the first specimen, a random number is drawn and compared to the probability of failure associated with the initial stress level. The stress level is increased or decreased depending on the result of this draw (if the random number is less than the probability of failure, the result is considered a failure, otherwise it is a survival). Another random number is then drawn for the second specimen and compared to the probability of failure for its associated stress level, with the stress level increased or decreased based on this comparison, and the process is repeated until all specimens are used. Note that test data must be bounded by both a stress level with zero probability of failure and unity probability or the staircase may walk to a stress level where no data exists. Through this algorithm, a set of virtual staircase tests is generated for the one real staircase test. For each virtual staircase, the Dixon and Mood method can be applied to provide a standard deviation estimate. The result is a distribution of standard deviation estimates. Lastly, the revised point estimate can then be taken as the mean of this distribution, or another statistic such as the median or other percentile point. The staircase simulation was modified to accommodate this bootstrapping algorithm by simulating a “real” staircase using an assumed underlying fatigue strength distribution, and then using the data from this simulated staircase to bootstrap additional staircases from which a distribution of standard deviation estimates can be analyzed.

Several comparison investigations were conducted between Dixon and Mood standard deviation estimates and bootstrapped estimates in order to determine the effectiveness of the bootstrap algorithm. For each case, the bootstrap algorithm was found to reduce the scatter in staircase results. The bootstrap had the effect of driving standard deviation estimates closer to their expected value given the step size and number of specimens. A sample comparison of bootstrap results to Dixon and Mood results was the case of a 12-specimen test conducted at a step size of 1 σ. In this case, 50 staircases were simulated, resulting in 50 standard deviation estimates (due to the discretized nature of staircase results, only 10 unique estimates resulted). For each of the 50 sets of staircase data, 5000 bootstrap replications were conducted with the mean of these 5000 points used as the bootstrap standard deviation estimate. Fig. 8 illustrates this data. Fig. 9 displays this same data in a histogram format.

The results of the bootstrap analysis show that the bootstrapping algorithm adjusted the estimate for standard deviation to be closer to the expected value given the test parameters in all cases. As Fig. 9 shows, the further the standard deviation estimate based on Dixon and Mood was away from the true standard deviation, the more the bootstrap was effective in closing the gap between the expected

![Comparison of normalized standard deviations for Dixon and Mood (mean), Svensson–Loren correction, and proposed correction for N = 20 specimens.](image-url)
value and the calculated value. Thus, the bootstrap provides a means of protection against large errors associated with small-sample tests due to statistical variance. Note, however, that standard deviation

![Figure 8](image_url)

**Fig. 8.** Bootstrap results sorted by Dixon and Mood standard deviation for 50 staircases with $N = 12$ and $s/\sigma = 1$.

![Figure 9](image_url)

**Fig. 9.** Frequency histograms for Dixon and Mood standard deviation and bootstrapped standard deviation for 50 staircases with $N = 12$ and $s/\sigma = 1$.

![Figure 10](image_url)

**Fig. 10.** Mean standard deviation estimates for 15-specimen staircase test with step $1\sigma$ resulting in four stress levels.
bias due to step size and sample size implies that the expected value of standard deviation for a staircase test is not the same as the true value. Use of a bias correction (Svensson–Loren or the proposed correction in this paper) in conjunction with the bootstrap helps to correct the expected value closer to the true value. In addition, analysis of various bootstrap statistics showed that use of the average of the 60th and 65th percentiles of the bootstrap distribution was more effective than use of the bootstrap mean for cases where the staircase resulted in four stress levels. For five or more stress levels, use of the bootstrap mean was more effective. When only two or three stress levels result, the bootstrap is not very effective, and staircase results should be used with extreme caution if they are to be used at all as the results tend to be highly dependent on step size alone. Figs. 10 and 11 compare the standard deviation results for 15 simulations resulting in 4-level results for a 15-specimen staircase with step 1\( \sigma \) result in four stress levels. Note that both the bias corrections were effective in reducing bias, but significantly increased scatter (larger standard deviation in the standard deviation estimates). However, when combined with bootstrapping, the results are less biased than Dixon and Mood but with similar or less scatter.

9. Conclusions

This simulation-based analysis has led to the quantification of the Dixon and Mood standard deviation bias as a function of both step size and sample size for staircase testing. It has been shown that Dixon and Mood equations generally underestimate standard deviation when traditional step sizes on the order of 2/3–3/2 of the true standard deviation are used. Step sizes on the order of 1.6–1.75 of the true standard deviation result in less bias and less scatter, but lead to staircase tests with fewer resulting stress levels which thereby leads to results too dependent on step size alone. Use of a proposed correction factor allows a larger range of less biased standard deviation estimation at smaller step sizes. This correction factor generally increases scatter in results, however. Incorporation of a bootstrapping algorithm was shown to significantly reduce scatter in standard deviation estimates, and in most cases reduces some bias as well. This simple, numerical procedure therefore provides a measure of protection against large errors when conducting small-sample staircase tests. Combination of a correction factor with bootstrapping allows less biased standard deviation estimates with little increase (and often a decrease) in statistical scatter. In addition, use of a starting stress several steps above or below the true fatigue strength mean acts to decrease standard deviation bias by increasing the probability of a staircase with more than one stress level with both failures and survivals. It is key in staircase analysis to have two or more stress levels with both failures and survivals in order for the bootstrapping algorithm to be effective. Thus, the combination of offset starting stress, a correction to the standard deviation equation, and bootstrapping provides a generally more reliable and robust means of calculating standard deviation from staircase data than traditional Dixon and Mood analysis, allowing possible use of the staircase algorithm for characterizing the fatigue strength distribution with fewer specimens.
References


