Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

EVALUATION OF TESTS FOR INITIAL-CONDITION BIAS

Charles R. Cash
Barry L. Nelson

Dept. of Industrial & Systems Engineering
The Ohio State University
Columbus, Ohio 43210, U.S.A.

David G. Dippold

American Electric Power Service Corporation
1 Riverside Plaza
Columbus, Ohio 43215, U.S.A.

J. Mark Long

ComputerPeople Consulting Services
50 Northwoods Boulevard
Worthington, Ohio 43225, U.S.A.

William P. Pollard

National Regulatory Research Institute
The Ohio State University
Columbus, Ohio 43210, U.S.A.

ABSTRACT

We evaluate, theoretically and empirically, the power of a family of tests for initial-condition bias.

1 INTRODUCTION

When the goal of a simulation experiment is to estimate steady-state parameters, the initial conditions of the simulation usually bias the estimators. This initial-condition-bias problem has been studied extensively.

Goldsman, Schruben, and Swain (GSS 1991) propose a family of tests for the presence of initial-condition bias that generalize and extend the tests proposed earlier by Schruben (1982) and Schruben, Singh and Tierney (1983). These tests are appealing because they are easy to implement and can be applied to the output of a single-replication experiment, which is the experiment design that is often recommended by researchers.

Using asymptotic results derived by GSS, and three simple models—the AR(1), the M/M/1 queue, and the Markov Chain—this paper attempts to answer several questions about these tests: When do the tests work and when do they fail? When the tests do work, which test is most powerful? And how does the batching strategy—which determines the degrees of freedom—affect the power of the tests?

2 THE TEST STATISTICS

Let $X_1, X_2, \ldots, X_n$ be a simulation output process in time-dependent order, and let $\bar{X}_n$ be the sample mean, a point estimator for the steady-state mean, μ. Using various functions of the original data, GSS derived a family of tests for initial-condition bias; that is, tests for the bias in $\bar{X}_n$ as an estimator of μ. All of the tests are based on an F statistic that compares the variability in the first portion of the output process to the variability in the latter portion of the process. The null hypothesis of *no initial-condition bias* is rejected if $F > \mathrm{F}_{1-\alpha,c,d}$, the $1-\alpha$ quantile of an F distribution with c and d degrees of freedom. The test statistics in GSS are reviewed in the following subsections.

2.1 Batch-Means Test

Partition $X_1, X_2, \ldots, X_n$ into b nonoverlapping batches of m observations such that $n = bm$, and define the following functions of the original data for $i = 1, 2, \ldots, b$:

$$\begin{aligned}
\bar{X}_i &= \frac{1}{m}\sum_{j=1}^{m} X_{(i-1)m+j} \\
Q_{BM} &= m\sum_{i=1}^{b}\left[\bar{X}_i - \frac{1}{b}\sum_{j=1}^{b}\bar{X}_j\right]^2 \\
V_{BM} &= \frac{Q_{BM}}{b-1}.
\end{aligned}$$

The quantities $\bar{X}_i$ and V_{BM} are the batch means and the batch-means variance estimator, respectively.

Suppose the b batch means are partitioned into two not necessarily equal-sized groups consisting of the first b' batch means and the last $b-b'$ batch means. Let V_{BM}^{1st} be the variance estimator defined above, but computed from only the first b' batches, and let V_{BM}^{2nd}

be the variance estimator defined above, but computed from only the last $b-b'$ batches. Under the null hypothesis, the ratio $F_{BM} = V_{BM}^{1st}/V_{BM}^{2nd}$ converges in distribution to an F random variable with appropriate degrees of freedom. The critical value for this test is $F_{1-\alpha,b'-1,b-b'-1}$.

2.2 Area Test

For the same b batches, transform the data into b standardized time series and compute a variance estimator based on the area under the standardized time series as follows, where $i=1,2,\ldots,b$, $j=1,2,\ldots,m$ and $0 \le t \le 1$:

$$\begin{aligned}
\bar{X}_{i,j} &= \frac{1}{j}\sum_{t=1}^{j} X_{(i-1)m+t} \\
T_{i,m}(t) &= \frac{\lfloor mt \rfloor(\bar{X}_{i,m} - \bar{X}_{i,\lfloor mt \rfloor})}{\sigma\sqrt{m}} \\
\hat{A}_i &= \frac{\sigma}{m}\sum_{j=1}^{m}\sqrt{12}\,T_{i,m}(j/m) \\
Q_{AREA} &= \sum_{i=1}^{b}\hat{A}_i^2 \\
V_{AREA} &= \frac{Q_{AREA}}{b}.
\end{aligned}$$

Here, $\lfloor\cdot\rfloor$ is the greatest integer function and σ^2 is the asymptotic variance constant of the output process.

Again suppose the b batches are divided into two parts. Let V_{AREA}^{1st} be the variance estimator from the first b' batches and let V_{AREA}^{2nd} be the variance estimator from the last $b-b'$ batches. Under the null hypothesis the ratio $F_A = V_{AREA}^{1st}/V_{AREA}^{2nd}$ converges in distribution to an F random variable with appropriate degrees of freedom. The critical value for this test is $F_{1-\alpha,b',b-b'}$.

2.3 Maximum Test

Using the cumulative within-batch means, $\bar{X}_{i,j}$, defined for the area test, a test based on the location and magnitude of the maximum deviation is possible. The test presented here is for the presence of negative bias; an analogous test is available for the presence of positive bias. For $i=1,2,\ldots,b$ and $j=1,2,\ldots,m$ compute

$$\begin{aligned}
S_{i,j} &= \bar{X}_{i,m} - \bar{X}_{i,j} \\
\hat{K}_i &= \operatorname{argmax}_{1\le j\le m}\{jS_{i,j}\} \\
\hat{S}_i &= \hat{K}_i S_{i,\hat{K}_i} \\
Q_{MAX} &= \sum_{i=1}^{b}\frac{m\hat{S}_i^2}{\hat{K}_i(m-\hat{K}_i)} \\
V_{MAX} &= \frac{Q_{MAX}}{3b}.
\end{aligned}$$

Again suppose the b batches are divided into two parts. Let V_{MAX}^{1st} be the variance estimator from the first b' batches and let V_{MAX}^{2nd} be the variance estimator from the last $b-b'$ batches. Under the null hypothesis the ratio $F_{MAX} = V_{MAX}^{1st}/V_{MAX}^{2nd}$ converges in distribution to an F random variable with appropriate degrees of freedom. The critical value for this test is $F_{1-\alpha,3b',3b-3b'}$.

2.4 Combined Tests

Because they are asymptotically independent, we can combine the batch-means test statistic with the area test statistic and with the maximum test statistic to create two more F-tests:

$$\begin{aligned}
Q_{BM+AREA} &= Q_{BM} + Q_{AREA} \\
V_{BM+AREA} &= \frac{Q_{BM+AREA}}{2b-1} \\
Q_{BM+MAX} &= Q_{BM} + Q_{MAX} \\
V_{BM+MAX} &= \frac{Q_{BM+MAX}}{4b-1}
\end{aligned}$$

Under the null hypothesis of no bias

$$F_{BM+AREA} = V_{BM+AREA}^{1st}/V_{BM+AREA}^{2nd}$$

and

$$F_{BM+MAX} = V_{BM+MAX}^{1st}/V_{BM+MAX}^{2nd}$$

are distributed as $F_{2b'-1,2b-2b'-1}$ and $F_{4b'-1,4b-4b'-1}$, respectively.

3 POWER CALCULATIONS

GSS derived expressions for the power of the BM, AREA, and BM+AREA tests under the following assumptions:

1. The output process $X_t \equiv Y_t - \mu a_t$, where $Y_1, Y_2, \ldots$ is a stationary process with mean μ, and $a_1, a_2, \ldots$ is a sequence of constants such that $a_t \longrightarrow 0$ as $t \longrightarrow \infty$. Thus, $\mathrm{E}[\bar{X}_n] = \mu - \bar{a}_n\mu$, and $\mathrm{Bias}[\bar{X}_n] = -\bar{a}_n\mu$, where

$$\bar{a}_n = \frac{1}{n}\sum_{t=1}^{n} a_t.$$

2. Statistics (batch means and squared area estimators, respectively) computed from different batches are independent and follow their asymptotic distributions (normal, χ^2, respectively).

For a fixed number of batches, $b = 20$, and a specific form of the transient bias, a_t (see (3) below), GSS studied the effect of changing b' ($b' = 5, 8$) on the power of the tests. They indexed their power calculations by μ^2/σ^2, the square of the steady-state mean in units of the limiting process variance.

In this section we use the expressions in GSS to examine the effect of the total number of batches, b, and the form of the bias, a_t, on the power. We index our power calculations by the *relative bias*

$$\sqrt{\frac{\mathrm{Bias}[\bar{X}_n]^2}{\sigma^2}} = |\bar{a}_n \sqrt{\frac{\mu^2}{\sigma^2}}$$

which is the bias of the sample mean in units of the limiting process variance. The relative bias is a measure of how significant the bias is relative to the noise in the process; power should be an increasing function of the relative bias. We constructed examples with different forms of the bias function, a_t, but the same relative bias; they are described below.

3.1 Bias Functions

We constructed four bias functions. In all cases $n = 2400$, $\mu/\sigma = 1$, and $\bar{a}_{2400} = r > 0$, so that the relative bias is r. In addition, and somewhat artificially, a_t goes to 0 at the midpoint of the output process; this allows us to examine the power of the tests when the assumptions are exactly satisfied and when they are not. The four bias functions are listed below:

Mean-shift bias function:

$$a_t = \begin{cases} 2r & t = 1, 2, \ldots, 1200 \\ 0 & t = 1201, \ldots, 2400 \end{cases} \tag{1}$$

Linear bias function:

$$a_t = \begin{cases} \frac{4800}{1199} r \left(1 - \frac{t}{1200}\right) & t = 1, 2, \ldots, 1200 \\ 0 & t = 1201, \ldots, 2400 \end{cases} \tag{2}$$

The constant $4800/1199 \approx 4$.

Quadratic bias function:

$$a_t = \begin{cases} \frac{1.728 \times 10^7}{2876401} r \left(1 - \frac{t}{1200}\right)^2 & t = 1, 2, \ldots, 1200 \\ 0 & t = 1201, \ldots, 2400 \end{cases} \tag{3}$$

The constant $1.728 \times 10^7/2876401 \approx 6$. GSS used this bias function, but with different constants.

Damped, oscillating bias function:

$$a_t = \begin{cases} \frac{1614r}{p} \sin\left(\frac{\pi t}{300}\right) & t = 1, 2, \ldots, 1200 \\ 0 & t = 1201, \ldots, 2400 \end{cases} \tag{4}$$

Figure 1 shows all four bias functions.

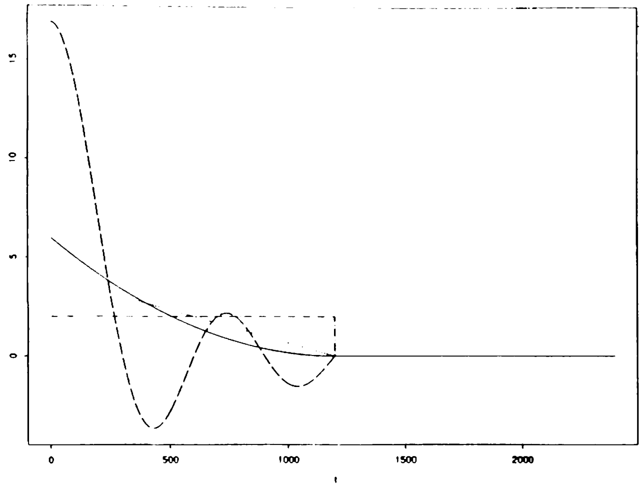

Figure 1: Bias Functions used in the Power Study

3.2 Results

We studied the power of the BM, AREA and BM+AREA tests against the bias processes (1)-(4) over the range of parameters given in Table 1. Notice the factor f, the fraction of the batches used to compute V^{1st}, which determines b'. For our bias functions, in which a_t goes to 0 at the midpoint of the output process, $f = 0.5$ should be the best choice. By varying f from 0.25 to 0.75 we examined the effect of dividing the process too early and too late, respectively.

Table 1: Cases Examined in Asymptotic Power Study

factor	parameters
total number of batches, b	2, 8, 16, 20, 40
relative bias, r	0.01, 0.1, 0.25
$f = b'/b$	0.25, 0.5, 0.75
confidence level, α	0.05
scale μ^2/σ^2	1

The power against relative bias $r = 0.1$ is shown in Tables 2-4. For all values of f, power increases as we move from shift to linear to quadratic to oscillating bias function. Qualitatively, the power increases the more pronounced the bias is at the beginning of the output process, even if it dies out quickly (recall that all four examples have the same relative bias of the sample mean). When $f = 0.5$, the tests have no power against the shift bias; this is not surprising since the tests are designed to detect differences between the variability in the first and second portions of the output process, but there is no difference when $f = 0.5$.

In most cases power is near its peak at $b = 16$ batches; in some cases it decreases beyond $b = 16$. In the next subsection we argue that we only need to consider small values of b.

Comparing Tables 2 and 4 to Table 3 shows that power is reduced by making b' either too large or too small. Of course, the practitioner never knows what "too large" or "too small" is.

Table 2: Power when Relative Bias $r = 0.1$ and $f = 0.25$

bias	b	BM	AREA	BM + AREA
shift				
	8	0.00	0.05	0.00
	16	0.00	0.05	0.00
	20	0.00	0.05	0.00
	40	0.00	0.05	0.01
linear				
	8	0.04	0.06	0.05
	16	0.05	0.05	0.05
	20	0.05	0.06	0.05
	40	0.05	0.05	0.05
quadratic				
	8	0.47	0.17	0.49
	16	0.47	0.07	0.41
	20	0.45	0.06	0.38
	40	0.38	0.05	0.29
oscillating				
	8	1.00	1.00	1.00
	16	1.00	0.86	1.00
	20	1.00	0.67	1.00
	40	1.00	0.16	1.00

The following is a summary of the results from the other cases we examined that are not presented here:

1. When $r = 0.01$, none of the tests showed power significantly larger than the size of the test, α.

2. When $r = 0.25$, the BM and BM+AREA tests had power nearly 1 for all but the shift bias process; the AREA test still had very low power.

3. The AREA test performed somewhat better against the linear and quadratic bias when $b = 2$ and $b' = 1$; it is the only test of the three that can be applied with just two batches. However, it is still inferior to the other two tests. Table 5 shows the power of the AREA test when $r = 0.1$ and $b = 2$.

Table 3: Power when Relative Bias $r = 0.1$ and $f = 0.5$

bias	b	BM	AREA	BM + AREA
shift				
	8	0.05	0.05	0.05
	16	0.05	0.05	0.05
	20	0.05	0.05	0.05
	40	0.05	0.05	0.05
linear				
	8	0.40	0.07	0.45
	16	0.44	0.05	0.39
	20	0.43	0.05	0.36
	40	0.36	0.05	0.28
quadratic				
	8	0.72	0.12	0.85
	16	0.84	0.06	0.83
	20	0.84	0.06	0.81
	40	0.80	0.05	0.70
oscillating				
	8	1.00	0.99	1.00
	16	1.00	0.66	1.00
	20	1.00	0.44	1.00
	40	1.00	0.11	1.00

In Section 4 we present small-sample, empirical estimates of the power of all five tests against more realistic examples.

3.3 Batch-Size Effects

To apply any of the tests we must choose b, the number of batches. For the BM test we will show that b should be relatively small, even if b could be made larger without violating the assumptions behind the test. We conjecture that the argument generalizes to the other tests.

For illustration, consider $f = 0.5$ (i.e., dividing the output process in half). Table 6 shows the F critical values for the BM test with $\alpha = 0.05$ at selected values of b. The minimum value of b is 4 for this test. The critical value initially decreases dramatically as b increases, but the marginal decrease diminishes rapidly. Remember that we reject the hypothesis of no bias if the test statistic exceeds the critical

Table 4: Power when Relative Bias $r = 0.1$ and $f = 0.75$

bias	b	BM	AREA	BM + AREA
shift				
	8	0.11	0.05	0.16
	16	0.16	0.05	0.17
	20	0.17	0.05	0.17
	40	0.17	0.05	0.14
linear				
	8	0.14	0.06	0.28
	16	0.28	0.05	0.33
	20	0.30	0.05	0.32
	40	0.32	0.05	0.28
quadratic				
	8	0.17	0.07	0.43
	16	0.42	0.06	0.55
	20	0.48	0.05	0.55
	40	0.55	0.05	0.50
oscillating				
	8	0.39	0.55	0.99
	16	0.99	0.28	1.00
	20	1.00	0.20	1.00
	40	1.00	0.08	1.00

Table 5: Power when Relative Bias $r = 0.1$ and $f = 0.5$

bias	b	AREA
shift	2	0.05
linear	2	0.25
quadratic	2	0.36
oscillating	2	0.65

value.

Of more importance is how the test statistic, F_{BM}, behaves relative to the critical value as a function of b. Suppose that we have an output process of length n, and that $a_t = 0$ for $t > n/2$ (i.e., there is no bias in the second half of the process). Define

$$\bar{a}_{i,m} = \frac{1}{m}\sum_{j=1}^{m} a_{(i-1)m+j}\,.$$

Let $f = 0.5$, so that $b' = b/2$. Using results in GSS we can show that $\mathrm{E}[F_{BM}]$ is equal to

$$\begin{aligned} & \left(\tfrac{b-2}{b-6}\right)\left(1+\tfrac{n\mu^2}{\sigma^2}\tfrac{1}{b'(b'-1)}\textstyle\sum_{i=1}^{b'}(\bar{a}_{i,m}-\bar{a}_n)^2\right) \\ = \quad & \left(\tfrac{b-2}{b-6}\right)\left(1+\tfrac{n\mu^2}{\sigma^2}\gamma(b/2)\right). \qquad (5) \end{aligned}$$

The term $(b-2)/(b-6)$ is the expected value of F_{BM} if there is no bias; $\gamma(b/2)$ represents the effect of the bias. Nelson (1990, Proposition 2) showed that $\gamma(b/2)$ will be a *decreasing* function of b when n is fixed for typical bias functions a_t. Therefore, we want to keep the number of batches small to make the test statistic significantly larger than its expectation when bias is present. Based on this result and the power calculations described above, we used $b \leq 16$ in the empirical study.

Table 6: Critical Values for the BM Test with $\alpha = 0.05$ and $b' = b/2$

b	$\mathrm{F}_{1-\alpha,b'-1,b'-1}$
4	161.
6	19.0
8	9.28
12	5.05
16	3.79
26	2.69
122	1.53
∞	1.00

4 EXPERIMENTS

We selected three models to study the power of the five tests: The AR(1) process, M/M/1 queue, and Markov chain. They are easy to simulate and known results allow us to calculate the bias and the variance of the process mean as a function of the process parameters and the length of the simulation. A disadvantage of these simple models is that it is difficult to simultaneously have a long run—a requirement for the tests to be valid—and significant bias; we discuss this issue below.

To evaluate the effectiveness of the tests against these different models, an index of the deviation from the null hypothesis is required. Define $\mathrm{Bias}_n = \mathrm{Bias}[\bar{X}_n]$, the bias of the point estimator at observation n, and $\sigma^2_{\bar{X}_n} = \sigma^2/n$, the asymptotic variance of the process divided by the number of observations, n; this is approximately $\mathrm{Var}[\bar{X}_n]$ in large samples. Then two candidate indices are:

$$\frac{|\ \mathrm{Bias}_n\ |}{\mu} \qquad (6)$$

$$\frac{|\ \mathrm{Bias}_n\ |}{\sigma_{\bar{X}_n}}. \qquad (7)$$

While index (6) has a natural interpretation and is similar to a measure used in GSS, index (7) can be thought of as a signal-to-noise ratio related to the length of the simulation, n. In our experiments we used index (7) and set this index to 0.1 and 0.25; that is, a relative bias of 10% or 25%, respectively, of the standard deviation of the point estimator $\bar{X}_n$. Certain model parameters and initial conditions will allow each of the three models to achieve this relative bias at a run length n that is large enough to allow batching. This in turn allows the application of the five tests defined above.

To study the implications of the batching strategy we apply five different b and b' combinations to the simulation output. Table 7 gives the batching strategies, defined as (b', b), exercised in the experiments. Notice that for BM and the combined tests the $(1, 2)$ strategy cannot be used.

Table 7: Batching Strategies (b', b)

b'	b
1	2
4	8
4	16
8	16
12	16

The experiments consisted of specifying the model parameters, initial conditions and run length n so that the relative bias was either 0.1 or 0.25. For each model thus specified, 1000 replications were generated. For each replication all five batching strategies were employed. And for each batching strategy, all five tests were applied, if possible, at the $\alpha = 0.05$ level. With this design, the power computed for each test (within batching strategy within model) is accurate to one digit after the decimal place with an additional digit useful for rounding.

In the subsections that follow we display results for relative bias 0.25. For some models the run length required to achieve this bias was quite short, leading to small batches. Thus, the effect of batch size (dependence between batches, convergence to asymptotic distributions within batches) are confounded with the properties of the individual tests to some extent. Nevertheless, we feel that it is important to maintain some fixed level of bias across examples to compare the results. We observed no improvement in the absolute performance or change in the relative performance of the tests at relative bias 0.1, which allowed significantly longer run lengths and batch sizes.

4.1 AR(1) Process

Let X_t be the tth term in the AR(1) process

$$X_t = \phi X_{t-1} + \varepsilon_t$$

for $t = 0, 1, \ldots n$, where the ε_t are i.i.d. $\mathrm{N}(0, (1-\phi)^2)$ random variables and the initial state x_0 is a constant. Formulas for Bias_n and the limiting process variance σ^2 are given by Kelton and Law (1984). We set $\phi = 0.7$ and 0.9, and varied x_0 and n to achieve the desired relative bias (7).

The marginal distribution of X_t is normal, which should be conducive to the tests. The $\mathrm{E}[X_t]$ converges monotonically to 0 for this process.

Table 8 shows the estimated power of the tests when $\phi = 0.9$ and $n = 1200$. For all batching strategies the power of the BM, AREA and BM+AREA tests is barely larger than the size of the test. The MAX and MAX+BM tests do better, particularly with $b = 16$ batches. For the best combination, $b' = 4, b = 16$, the relative bias at the break point ($n = 300$) is 0.50, twice as large as the relative bias for the entire process.

Table 8: Results for AR(1) with $\phi = 0.9$, Initial State $x_0 = -0.96$ and $n = 1200$, implying Relative Bias 0.25

		Estimated Power				
b'	b	BM	AREA	BM+A	MAX	BM+M
1	2		0.06		0.27	
4	8	0.07	0.08	0.08	0.47	0.43
4	16	0.10	0.17	0.19	0.79	0.76
8	16	0.06	0.12	0.11	0.61	0.59
12	16	0.06	0.08	0.08	0.42	0.38

4.2 M/M/1 Queue

Let X_t be the delay in queue of the tth customer arriving to an M/M/1 queue with arrival rate λ, service rate 1 and k customers present at time 0. The Bias_n can be calculated using recursive algorithms in Kelton and Law (1985), and the limiting process variance σ^2 is given by Whitt (1989). We set $\lambda = 0.5$ and 0.8 (implying traffic intensity $\rho = 0.5$ and 0.8), and varied k and n to achieve the desired relative bias (7).

The steady-state marginal distribution of X_t is a mixture of an exponential distribution and a point mass at 0. Depending upon the choice of k, $\mathrm{E}[X_t]$ converges to μ monotonically from below, from above, or crosses μ from above then converges monotonically from below.

Table 9 shows the estimated power of the tests when $\rho = 0.8$, $k = 10$ and $n = 112$; $\mathrm{E}[X_t]$ converges to μ monotonically from above for this example. The run length is quite small, implying very small batches.

The results are nearly the same when the relative bias is 0.1, which allows a run length of $n = 736$.

The BM test does somewhat better in this example relative to the AR(1) example. The BM+MAX test does the best at combination, $b' = 12, b = 16$; the relative bias at the break point ($n = 84$) is 0.27, so there must still be significant bias after the break point.

Table 9: Results for M/M/1 with $\rho = 0.8$, Initial Condition $k = 10$ and $n = 112$, implying Relative Bias 0.25

		Estimated Power				
b'	b	BM	AREA	BM+A	MAX	BM+M
1	2		0.11		0.47	
4	8	0.22	0.14	0.31	0.49	0.45
4	16	0.16	0.15	0.19	0.48	0.27
8	16	0.31	0.17	0.39	0.47	0.49
12	16	0.32	0.13	0.50	0.37	0.62

4.3 Markov Chain

Let X_t be the state after transition t of a finite-state, ergodic Markov chain with one-step transition matrix $\mathbf{P}$, where the initial state x_0 is fixed. Formulas for Bias_n and the limiting process variance σ^2 are given by Glynn (1984). We constructed two transition matrices and varied x_0 and n to achieve the desired relative bias (7). The two transition matrices are shown below: $\mathbf{P}_1$ is a five-state chain with state space $\{1, 2, 3, 4, 5\}$, while $\mathbf{P}_2$ is a ten-state chain with state space $\{1, 2, \ldots, 10\}$.

$$\mathbf{P}_1 = \begin{bmatrix} 0.2 & 0 & 0.6 & 0.2 & 0 \\ 0 & 0.3 & 0 & 0 & 0.7 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0.4 & 0 & 0.6 & 0 \\ 0.1 & 0 & 0 & 0.3 & 0.6 \end{bmatrix}$$

$$\mathbf{P}_2 =$$

$$\begin{bmatrix} 0.99 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.3 & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.9 \end{bmatrix}$$

The marginal distribution of X_t is discrete. The five-state chain is relatively balanced, while the ten-state chain is relatively unbalanced in the sense that it has two nearly absorbing states which are distant from each other. Consequently, one would expect the extent and duration of the bias to be strongly dependent on the initial state for $\mathbf{P}_2$. The intent of the unbalance is to allow us to induce a large bias. For both chains $\text{E}[X_t]$ converges to μ monotonically.

Table 10 shows the estimated power of the tests for $\mathbf{P}_1$ with initial state 1 and $n = 160$. The run length is quite small in this example, implying very small batches. Similar to the M/M/1 example, the power is low, barely larger than the size of the test for the BM, AREA and BM+AREA tests.

Table 11 shows the estimated power of the tests for $\mathbf{P}_2$ with initial state 5 and $n = 928$. The run length is much longer here, but with the same relative bias as the previous example. The power of the tests improves dramatically. At the midpoint of the process ($n = 464$), the relative bias is 0.35, quite a bit larger than the relative bias for the entire process.

Table 10: Results for $\mathbf{P}_1$ with Initial State $x_0 = 1$ and $n = 160$, implying Relative Bias 0.25

		Estimated Power				
b'	b	BM	AREA	BM+A	MAX	BM+M
1	2		0.06		0.20	
4	8	0.05	0.07	0.07	0.36	0.32
4	16	0.08	0.12	0.11	0.57	0.52
8	16	0.07	0.07	0.07	0.40	0.34
12	16	0.06	0.07	0.06	0.21	0.16

Table 11: Results for $\mathbf{P}_2$ with Initial State $x_0 = 5$ and $n = 928$, implying Relative Bias 0.25

		Estimated Power				
b'	b	BM	AREA	BM+A	MAX	BM+M
1	2		0.22		0.62	
4	8	0.32	0.48	0.49	0.73	0.72
4	16	0.37	0.51	0.50	0.68	0.68
8	16	0.43	0.54	0.54	0.73	0.72
12	16	0.50	0.59	0.61	0.76	0.75

4.4 Size of the Tests

Subsections 4.1–4.3 examined the power of the tests when bias is present. Also important is the *size* of each test, which is the probability that a test rejects the null hypothesis of no bias when there is in fact no bias. Asymptotically, the size of all of the tests is α. In this subsection we estimate the size of the tests in small samples.

We repeated all of the experiments in Subsections 4.1–4.3 with two additional sets of initial conditions: Starting each process from the minimum-bias initial state (e.g., $x_0 = 0$ for the AR(1) model), and randomly sampling the initial state from the steady-state distribution. From these experiments we estimated the size of each test. The results are summarized in the following paragraph and representative results for the AR(1) process are presented.

In all cases the estimated size of the tests was less than the estimated power of the tests presented in Subsections 4.1–4.3. However, the estimated size was frequently larger than the nominal size, $\alpha = 0.05$, particularly in the examples with short run lengths. As b decreased (fewer, larger batches), the estimated size decreased toward the nominal size, as expected. The convergence was slowest for the MAX and BM+MAX tests.

Table 12 shows the results for the AR(1) model initialized at $x_0 = 0$, for which $\text{Bias}_n = 0$. These results are representative of what we observed for the other models and for random initialization. We found that the size of the MAX test converged to the nominal level at a run length of about $n = 4992$ and $b = 2$ batches for this model.

Table 12: Results for AR(1) with $\phi = 0.9$, Initial State $x_0 = 0$ and $n = 1200$, implying Relative Bias 0

		Estimated Size				
b'	b	BM	AREA	BM+A	MAX	BM+M
1	2		0.04		0.18	
4	8	0.05	0.05	0.05	0.27	0.24
4	16	0.05	0.05	0.05	0.42	0.38
8	16	0.05	0.05	0.06	0.32	0.25
12	16	0.06	0.07	0.06	0.19	0.16

5 DISCUSSION

In the introduction we posed three questions regarding the tests for initial-condition bias: When do the tests work and when do they fail? When the tests do work, which test is most powerful? And how does the batching strategy—which determines the degrees of freedom—affect the power of the tests? We offer some answers here.

Recall that in the examples we examined in Sections 3 and 4 we maintained a fixed relative bias of the point estimator, $\bar{X}_n$, for different bias functions $\mathrm{E}[X_t - \mu]$. The tests seem to be most powerful when the bias is severe at the very beginning of the output process, but dies out quickly. The more slowly the bias decays, the more difficulty the tests have detecting it.

The MAX and MAX+BM tests were not included in the asymptotic analysis because their power functions are intractable. Fortunately, our small-sample study gives conclusive evidence regarding the relative power of the tests: The MAX test is the most powerful, while the BM and AREA tests are the least powerful. Unfortunately, the power of the MAX may test come at the expense of a larger size than the nominal α level. The reader should be cautious regarding the actual numerical values we reported. Many of the experiments involved small batch sizes, so the performance of the tests may have been influenced by the dependence between batches and the distributions of the within-batch quantities. We note that in those cases that allowed longer runs and larger batch sizes the results were not markedly improved.

In addition, the tests all test for *statistically* significant bias, when, of course, we are most interested in whether there is *practically* significant bias. The small-sample results we displayed are for a bias that is 25% of the variability of the point estimator; whether or not that is practically significant depends on the application.

The batching strategy, which includes the total number of batches, b, and the fraction of those batches allocated to the numerator of the test, f, clearly affects the power of the tests. Results from the asymptotic study indicate that there is little if any benefit from making b excessively large, even if one can do so without jeopardizing the assumptions behind the tests. On the other hand, having b small helps insure that the assumptions behind the tests are valid, although if b is too small then power is lost. Based on the results presented here and others not reported, we recommend the MAX test with $b = 8$ batches, provided the run length is long enough so that each of the 8 batches is large. This value strikes a balance between obtaining high power when there is bias, and maintaining the desired size when there is no bias.

Choosing a value of f is more difficult. Ideally, we should divide the process at the point where the bias becomes negligible. Since that point is never known, $f = 0.75$ could be used to increase the chance that there is little bias in the second portion of the output process. Unfortunately, if the hypothesis of no bias is rejected and the remedy for bias is data deletion, then it is not clear how much data to delete. Rejecting the hypothesis when $f = 0.75$ does not mean that 75% of the data must be discarded.

The deletion strategy that should be used in conjunction with the tests is still an open problem. We have discussed several strategies, including the following: First perform the test with $f = 0.25$; if the null hypothesis is rejected, delete the first 25% of the data and apply the test again to the remaining data. If the null hypothesis is accepted, retest at $f = 0.5$ (and next at $f = 0.75$); the retest is needed because we may accept the hypothesis when there is significant bias in both the first and second portion of the process.

ACKNOWLEDGEMENTS

The authors acknowledge helpful discussions with Dave Goldsman. Barry Nelson's work was partially supported by NSF Grant DDM-8922721.

REFERENCES

Glynn, P. W. 1984. Some asymptotic formulas for Markov chains with applications to simulation. *Journal of Statistical Computation and Simulation* 19: 97–112.

Goldsman, D., L. Schruben and J. J. Swain. 1991. Tests for transient means in simulated time series. Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Kelton, W. D., and A. M. Law. 1984. An analytical evaluation of alternative strategies in steady-state simulation. *Operations Research* 32: 169–184.

Kelton, W. D., and A. M. Law. 1985. The transient behavior of the M/M/s queue, with implications for steady state-simulation. *Operations Research* 33: 378-396.

Nelson, B. L. 1990. Variance reduction in the presence of initial-condition bias. *IIE Transactions* 22: 340–350.

Schruben, L. 1982. Detecting initialization bias in simulation output. *Operations Research* 30: 569–590.

Schruben, L., H. Singh and L. Tierney. 1983. Optimal tests for initialization bias in simulation output. *Operations Research* 31: 1167–1178.

Whitt, W. 1989. Planning queueing simulations. *Management Science* 35: 1341–1366.

AUTHOR BIOGRAPHIES

CHARLES R. CASH is a Ph.D. student in the Department of Industrial and Systems Engineering at The Ohio State University. His research interests include simulation methodology, stochastic processes, and analysis of manufacturing systems. He is a member of IIE, SCS and ORSA.

DAVID G. DIPPOLD is a Senior Analyst at the American Electric Power Service Corporation, and a Ph.D. student in the Department of Industrial and Systems Engineering at The Ohio State University. His research interests are in the areas of simulation and optimization of large engineering systems.

J. MARK LONG is a Ph.D. student in the Department of Industrial and Systems Engineering at The Ohio State University, and is a consultant for Computer People Consulting Services, Columbus, Ohio. His research interests are simulation and parallel computing. He is a member of ORSA and ACM.

BARRY L. NELSON is an Associate Professor in the Department of Industrial and Systems Engineering at The Ohio State University. His research interests are experiment design and analysis of stochastic simulations. He is President of the TIMS College on Simulation and an Associate Editor for *Operations Research.*

WILLIAM P. POLLARD is a Senior Institute Engineering Economist at the National Regulatory Research Institute at The Ohio State University. He is also a Ph.D. student in the Department of Industrial and Systems Engineering at Ohio State. His research interests are telecommunications cost modelling, manufacturing systems engineering and simulation of stochastic systems.