INITIAL CONDITION BIAS AND EXPERIMENTAL DESIGN IN QUEUING SIMULATIONS

Mark A. Turnquist
Northwestern University, Evanston, Illinois

Joseph M. Sussman
Massachusetts Institute of Technology, Cambridge, Massachusetts

INTRODUCTION

The simulation of complex queuing systems is an important area of application for discrete event digital simulation. In many cases, the object of the analysis is the estimation of "steady-state" measures for the system. Since a simulation run does not typically begin with the system in a steady-state condition, the analyst must be concerned with the effects of the initial conditions in the system on the data being collected. For systems with a high traffic intensity, the time required for the simulation results to converge to steady-state may be very long indeed, as has been discussed by Eilon and Chowdhury (1).

It should be noted that this analysis applies to situations in which there is neither a well-defined "end" to a simulation run nor system-defined initial conditions. Thus, systems whose operations follow a natural cycle (e.g., a business which opens at 9:00 a.m. and closes at 5:00 p.m.) are not included. For such situations, the questions of initial conditions and run length are defined by the system under study, and are not at the discretion of the analyst in the same sense as in systems which have no natural cycle.

General recognition of the possible "bias" introduced in experimental data as a result of initial conditions has prompted several suggestions of methods to reduce this influence. (See, for example, Conway (2) and Fishman (3).) One common approach is to specify a "warm-up" period at the beginning of the run, during which no data are collected. Discarding this initial data typically does result in a bias reduction, although the exact amount of bias reduction achieved by warming up a given model for a particular length of time is usually unknown. Moreover, the discarding of data results in an increase in the variance of estimates of quantities of interest, such as mean queue length, mean waiting time, etc. The major issue in deciding on the length of this warm-up period is thus the trade-off between reduction in initial condition bias and increase in variance of statistical estimates, given that the analyst is operating with a fixed experimental budget.

A model with which to investigate the effects of initial conditions and warm-up period has been advanced by Fishman (4), based on autoregressive processes. The work in this paper extends that model to explore more completely the experimental design implications. There is also discussion of how the sensitive parameters of that model are influenced by traffic intensity in the systems under study, in order to provide some basic rules of thumb for simulation analysts in constructing experiments.

THE EXPERIMENTAL DESIGN PROBLEM

The experimental design problem is formulated in the following way. Under the premise that the simulation of one customer is the basic unit of cost in the simulation of queuing systems, a budget constraint of the form "total customers simulated ≤ M" is imposed.

The objective of the analyst is to estimate, with maximum effectiveness, some steady-state quantity of interest (e.g., expected wait time), subject to the budget constraint. Two of the most important measures of effectiveness for an estimator are bias and variance. Bias measures systematic deviation from the true mean, while variance measures variation about the bias plus the mean, as illustrated in Figure 1. As a single figure of merit for this discussion, we will use the mean-square error:

$$\text{MSE} = \text{variance} + (\text{bias})^2.$$

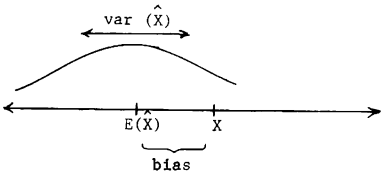

Figure 1. Composition of mean-square error for an estimator, $\hat{X}$, of a quantity, X.

In the attempt to design experiments so as to minimize MSE for an estimate of the quantity of interest, the analyst will be assumed to have three basic degrees of freedom:

- N = the number of independent replications of a simulation run
- W = "warm-up" period (number of customers) for any given simulation run
- L = length of data-collection period (number of customers) for any given simulation run.

The length of any given simulation is W+L, and the total effort involved in the experimental program is then N(W+L).

Hence, writing MSE as a function of N, W, and L, the experimental design problem can be formulated as follows:

$$\begin{aligned} &\min \text{MSE}(N,W,L) \\ &\text{s.t. } N(W+L) \leq M \\ &\quad N, W, L \text{ integer.} \end{aligned}$$

An understanding of the exact way in which MSE for estimates of steady-state queue length and waiting time depends upon N, W, and L requires first, an assessment of bias and variance given an initial condition; and second, a model which allows analysis of the effect of warm-up periods of various lengths on the system state at the beginning of the data-collection period. These requirements may be satisfied through use of an autoregressive model.

BIAS AND VARIANCE OF ESTIMATORS IN AN AUTOREGRESSIVE MODEL

A previous paper by Fishman (4) has explored the impacts of initial conditions on estimates in first-order autoregressive models. Study of such models is useful because their bias behavior is similar to that of many queueing models. The basic autoregressive model may be described briefly as follows.

Suppose that $\{X_t;\ t=0,\ldots\infty\}$ represents a process of interest in a simulation experiment, and that observations are taken at time $t=1,2,\ldots,n$. If X_0 is the initial state of the process, $\{X_t\}$ is said to be a first-order autoregressive process if it satisfies the following recursive expression:

$$[X_t - E(X)] = \alpha[X_{t-1} - E(X)] + \epsilon_t \qquad (1)$$

with

$$0 < \alpha < 1$$

$$E(\epsilon_t) = 0$$

$$E(\epsilon_s \epsilon_t) = \begin{cases} \sigma^2 & s = t \\ 0 & s \neq t \end{cases}$$

The conditional mean of $\{X_t\}$, given X_0, may be written as shown in (2).

$$E(X_t|X_0) = X_0\alpha^t + E(X)[1-\alpha^t] \quad (2)$$

Note that the dependence on X_0 declines geometrically.

The problem in which we are interested is the estimation of E(X), the steady-state expected value of the process. Suppose E(X) is estimated by

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t \quad (3)$$

Then the following result is obtained:

$$E(\bar{X}_n|X_0) = \frac{X_0\alpha(1-\alpha^n)}{n(1-\alpha)} + E(X)\left[1 - \frac{\alpha(1-\alpha^n)}{n(1-\alpha)}\right] \quad (4)$$

The bias in $\bar{X}_n$ is then easily shown to be as follows:

$$E(\bar{X}_n|X_0) - E(X) = \frac{[X_0 - E(X)]\alpha(1-\alpha^n)}{n(1-\alpha)} \quad (5)$$

Note that the bias is a function of:

$[X_0 - E(X)]$, the deviation of the initial state from the mean;
α, the damping constant;
and
n, the sample size.

Clearly, the best value for X_0 is E(X); but this value, of course, is unknown.

The variance of $\bar{X}_n$ is as shown in equation (6).

$$V(\bar{X}_n|X_0) = \frac{\sigma^2}{n(1-\alpha)^2}\left[1 - \frac{2\alpha(1-\alpha^n)}{n(1-\alpha^2)} - \frac{\alpha^2(1-\alpha^n)^2}{n(1-\alpha^2)}\right] \quad (6)$$

Note that the variance of $\bar{X}_n$ is <u>not</u> a function of the initial condition, X_0. It depends only on α and n.

Combining equations (5) and (6), the mean-square error of $\bar{X}_n$ can be constructed, as shown in equation (7).

$$MSE(\bar{X}_n|X_0) = \frac{[X_0-E(X)]^2\alpha^2(1-\alpha^n)^2}{n^2(1-\alpha)^2} + \frac{\sigma^2}{n(1-\alpha)^2}\left[1 - \frac{2\alpha(1-\alpha^n)}{n(1-\alpha^2)} - \frac{\alpha^2(1-\alpha^n)^2}{n(1-\alpha^2)}\right] \quad (7)$$

We are now in a position to consider the effect of discarding the first W observations from the sample, retaining only the last L (where W+L = n, in our previous notation). The estimator of E(X) is now

$$\bar{X}_{W,L} = \frac{1}{L}\sum_{t=W+1}^{L+W} X_t \quad (8)$$

The following relations hold:

$$E(\bar{X}_{W,L}|X_0) - E(X) = \frac{[X_0-E(X)]\alpha(\alpha^W-\alpha^{W+L})}{L(1-\alpha)} \quad (9)$$

$$V(\bar{X}_{W,L}|X_0) = \frac{\sigma^2}{L(1-\alpha)^2}\left\{1 - \frac{\alpha(1-\alpha^L)}{L(1-\alpha^2)}\left[2 + \alpha^{2W+1}(1-\alpha^L)\right]\right\} \quad (10)$$

The mean-square error of $\bar{X}_{W,L}$ is formed by combining equations (9) and (10), as shown in equation (11).

$$MSE(\bar{X}_{W,L}|X_0) = \frac{[X_0-E(X)]^2\alpha^2(\alpha^W-\alpha^{W+L})^2}{L^2(1-\alpha)^2} + \frac{\sigma^2}{L(1-\alpha)^2}\left\{1 - \frac{\alpha(1-\alpha^L)}{L(1-\alpha^2)}\left[2 + \alpha^{2W+1}(1-\alpha^L)\right]\right\} \quad (11)$$

The final element, that of considering N independent replications of a shorter simulation run, rather than a single long run, can now be introduced. Replication of the simulation runs, given that they all begin with the same initial condition, X_0, has no effect on bias. It does, however, affect the variance term. If the variance of $(\bar{X}_{W,L}|X_0)$ for a single run is given by equation (10), the variance of a sample mean based on N independent replications of such a run is given by

$$V(\bar{X}_{N,W,L}|X_0) = V(\bar{X}_{W,L}|X_0)/N \quad (12)$$

where
$$\bar{X}_{N,W,L}|X_0 = \frac{1}{N}\sum_{i=1}^{N}(\bar{X}^i_{W,L}|X_0) \quad (13)$$

and
$\bar{X}^i_{W,L}$ = sample mean from run i.

This results in the complete expression for mean-square error, as shown in equation (14).

$$MSE(\bar{X}_{N,W,L}|X_0) = \frac{[X_0-E(X)]^2\alpha^2(\alpha^W-\alpha^{W+L})^2}{L^2(1-\alpha)^2} + \frac{\sigma^2}{LN(1-\alpha)^2}\left\{1 - \frac{\alpha(1-\alpha^L)}{L(1-\alpha^2)}\left[2+\alpha^{2W+1}(1-\alpha^L)\right]\right\} \quad (14)$$

The experimental design problem is thus to find values of N, W, and L so as to

$$\begin{aligned} &\text{minimize } MSE(\bar{X}_{N,W,L}|X_0) \\ &\text{s.t.} \quad N(W+L) \le M \\ &\qquad N, W, L \text{ integer.} \end{aligned} \quad (15)$$

Given values of

$[X_0-E(X)]$, the initial departure from steady-state;
α, the damping constant; and
σ^2, the variance of the process $\{X_t\}$,

the solution to this constrained optimization problem yields the optimal experimental program.

<u>OPTIMAL EXPERIMENTS</u>

While a closed-form analytic solution to the optimization problem in (15) has not been obtained, some general hypotheses can be advanced about the nature of the optimal solution as a function of the key parameters, $[X_0-E(X)]$, α, and σ^2:

1) As $[X_0-E(X)]\to 0$, the bias term disappears and MSE is composed entirely of variance. In this case, it becomes advantageous to make many short runs with no warm-up. That is, N increases, L decreases, and $W\to 0$.

2) Conversely, as σ^2 decreases relative to $[X_0-E(X)]$, the bias term becomes more important, and one expects that the tendency is to make fewer runs, with longer warm-up periods and run lengths. That is, N decreases, and L and W increase.

3) As α increases, indicating a higher degree of autocorrelation in the observations, the influence of initial conditions decays very slowly and bias

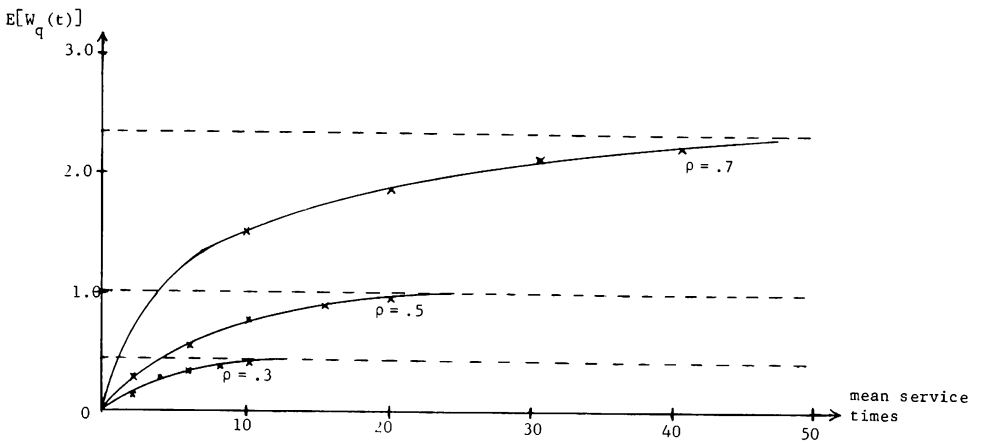

Figure 2. Transient behavior of expected wait time for several values of ρ, assuming "empty and idle" initial conditions.

becomes relatively more important than variance. The expected result is a tendency to make runs with longer warm-up.

In general, the solution to the experimental design optimization problem will be found by inputting values for $[X_0-E(X)]$, α, and σ^2, and applying direct search techniques to the minimization of MSE. The remainder of this paper will utilize this approach to explore the design of experiments for the simple illustrative case of an M/M/1 queuing system. Since both transient and steady-state solutions for this simple system are well known, it would not be necessary to simulate such a system in practice. Because of its simplicity, however, it does provide a useful test case, and may well provide important insights for more complex queuing systems.

AN ILLUSTRATIVE EXAMPLE

The primary focus in this analysis will be on the situation in which the system starts "empty and idle." This case is an important one because the empty and idle initial condition is perhaps the most frequently employed in practice. A thorough analysis of this case also provides a valuable benchmark against which to assess the effects of using other sets of initial conditions.

For the purposes of illustration, we will limit attention to mean waiting time of customers as the measure of interest in the system. While other measures will often be of interest as well, this provides a useful example.

In order to investigate optimal experiments for the M/M/1 model, we must first develop values for $[X_0-E(X)]$, α, and σ^2. The following sections discuss this process in some depth, as a basis for rules-of-thumb applicable to more complex queuing models.

Initial Deviation from Steady-State

Since the process of interest, $\{X_t\}$, is the waiting time of customers in the queue, an empty and idle initial state implies $X_0=0$. The value of $[X_0-E(X)]$ is thus simply $-E(X)$, or the negative of the expected wait time. In the simple M/M/1 queue, the solution for expected wait time is well known:

$$E(X) = \frac{\lambda}{\mu(\mu-\lambda)} \qquad (16)$$

where λ = mean arrival rate
μ = mean service rate.

In more complicated systems, of course, this quantity will not be known (and indeed, the object of the simulation is the estimation of this quantity.) However, it may be possible to construct a crude estimate on the basis of simple analysis which may be used for this procedure.

Variance of Wait Time

Again, because of the simplicity of the M/M/1 example, the variance of wait time is a known quantity:

$$\sigma^2 = \frac{\lambda}{\mu(\mu-\lambda)^2} \qquad (17)$$

While this quantity will also typically be unknown, it is often possible to derive at least an approximation to the probability distribution of the number of customers in the system, from which an estimate of σ^2 can be derived.

Damping Constant, α

The specification of α is of considerable importance to the experimental design process, and thus it is of interest to construct a general model which will provide guidance to the analyst in this regard. A model has been developed for the M/M/1 queue for the case in which the system starts empty and idle, and work is currently underway to generalize this model to include other initial conditions. The model is based on an analysis of the transient solution for M/M/1 systems.

The transient behavior of M/M/1 queuing systems has been discussed by a number of previous authors. A good discussion of one derivation is given by Morse (5). His basic approach is to analyze a system with a finite capacity, say Q. (The assumption is thus made that if a customer arrives when there are Q customers already in the system, i.e., a queue length of Q-1, he "balks" and leaves the system without being served.) The solution for the infinite-queue case is then obtained through a limiting process.

For an M/M/1 system with maximum capacity Q, Morse (5) has shown that if there are k_0 customers in the system (including the customer being served, if any) at time t=0, the probability of there being exactly k customers in the system at time $t=\tau$ is:

$$p_k(\tau) = P_k + \frac{2\rho^{\frac{1}{2}(k-k_0)}}{Q+1} \sum_{s=1}^{Q} \left(\frac{1}{\theta_s}\right)\left[\sin\left(\frac{sk_0\pi}{Q+1}\right) - \sqrt{\rho}\sin\left(\frac{s(k_0+1)\pi}{Q+1}\right)\right]$$

$$\cdot\left[\sin\left(\frac{sk\pi}{Q+1}\right) - \sqrt{\rho}\sin\left(\frac{s(k+1)\pi}{Q+1}\right)\right]e^{-\gamma_s\tau} \qquad (18)$$

where $p_k(\tau)$ = probability of observing k customers in the system at time $t = \tau$

P_k = steady-state probability of finding k customers in the system

$$= \left[\frac{1-\rho}{1-\rho^{Q+1}}\right]\rho^k$$

ρ = λ/μ (utilization rate of server)

λ = mean arrival rate of customers

μ = mean service rate

$$\theta_s = \frac{\gamma_s}{\mu}$$

$$\gamma_s = \lambda + \mu - 2\sqrt{\lambda\mu}\cos\left(\frac{s\pi}{Q+1}\right) \quad s = 1,2,\ldots,Q.$$

Given the set of $p_k(t)$, $k = 0,1,2,\ldots,Q$, the computation of expected queue length at time t, denoted $E[L_q(t)]$, is very straightforward, as shown in equation (19):

$$E[L_q(t)] = \sum_{k=1}^{Q-1} k p_{k+1}(t) \qquad (19)$$

Expected waiting time for a customer arriving at time t, denoted $E[W_q(t)]$, can also be determined easily, through use of a basic theorem which results in equation (20).

$$E[W_q(t)] = \frac{E[L_q(t)]}{\lambda} \qquad (20)$$

where λ is the mean arrival rate as defined previously.

If we now consider sampling the process $E[W_q(t)]$, defined by equations (18), (19) and (20), at intervals corresponding to one mean service time, an autoregressive model corresponding to equation (1) can be developed.

Figure 2 illustrates the mean waiting time computed from (18), (19) and (20), as well as the fitted autoregressive models, for several values of ρ, the traffic intensity. Column 2 of Table 1 shows the estimated values of α, computed from linear regression, for the various values of ρ.

(1) ρ	(2) $\hat{\alpha}$ (from regression)	(3) $\tilde{\alpha} = e^{-(1-\rho)^2}$
.1	.440	.445
.3	.606	.613
.5	.779	.779
.7	.896	.914
.9	.975	.990

Table 1. Estimated damping coefficients for several values of traffic intensity.

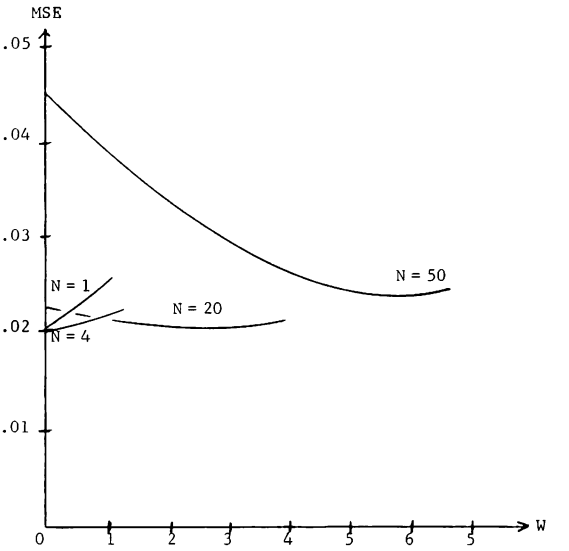

Figure 3a. MSE as function of W for several values of N, at $\rho = .5$

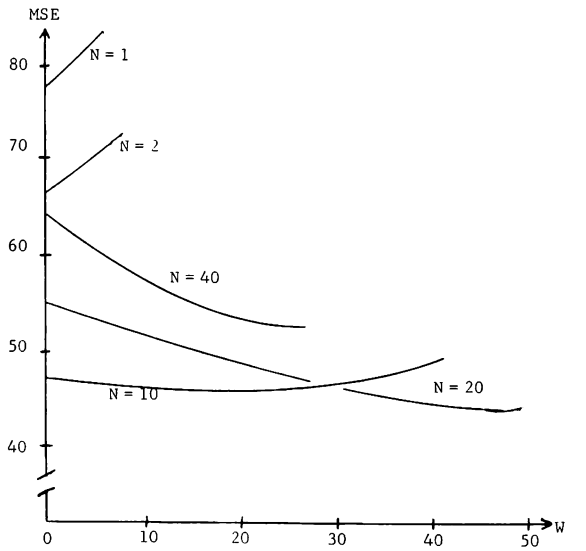

Figure 3b. MSE as function of W for several values of N at $\rho = .9$

Column 3 of Table 1 shows estimates of α computed from the function

$$\tilde{\alpha} = e^{-(1-\rho)^2} \qquad (21)$$

Comparison of Columns 2 and 3 of Table 1 indicates that a very reasonable approximation for the autoregressive damping constant may be obtained from the function in (21). Thus, knowing only the traffic intensity, ρ, for the system under study, the analyst can construct an estimate of α.

In summary, then, the mean-square error (to be minimized) can be expressed by substituting the results from equations (16), (17) and (21) into equation (14).

Experimental Design

Given the values of $[X_0-E(X)]$, α, and σ^2, a direct search procedure may be used to determine optimal experiments. As an example of the procedure, this has been done with the budget constraint M = 1000 customers, for values of ρ of .5 and .9.

The behavior of MSE as a function of W, the warm-up period, for several values of N, the number of replications, is shown in Figure 3. Figure 3a is the case when $\rho = .5$, and Figure 3b when $\rho = .9$.

Several important points are illustrated by Figure 3. The first of these is the relative insensitivity of MSE to N for $\rho = .5$. It is clearly more important in this case to allocate correctly between W and L, given N, than to choose N optimally. To some extent, the insensitivity observed is due to the rather large value of M, relative to what is required to obtain good estimates for a system with medium traffic intensity, such as this.

It is also interesting to note that, in the case of $\rho = .5$, the commonly-used procedure of one long run with no warm-up (i.e., N = 1, W = 0, L = 1000) is a nearly-optimal choice.

In the case where $\rho = .9$, the choice of N is more important, with a difference of about a factor of 2 in MSE between the optimal choice (N = 20) and the choice N = 1, for example. However, even in this case, over the range of about N = 10 to N = 20, MSE is quite insensitive to the exact choice.

Note also the drastically different ratio of W to L in the cases $\rho = .5$ and $\rho = .9$. The optimal policy for $\rho = .5$ is N = 4, W = 0, L = 250. For $\rho = .9$, it is N = 20, W = 49, L = 1. On an intuitive basis, as ρ increases (increasing α), the influence of initial conditions becomes more persistent and the data stream collected becomes more correlated. As a result, it pays to let the model warm up to reduce the initial condition bias, and then collect a very short sample record, which is replicated several times. In interpreting this result, it is important to point out that the choice of the "empty and idle" initial condition is clearly sub-optimal, and that other results might be obtained with more realistic initial conditions.

CONCLUSIONS

The results illustrated in Figure 3 are evidence of the elusiveness of really general rules-of-thumb with regard to experimental design. Furthermore, the preliminary results obtained thus far provide a somewhat tenuous basis for generalization. However, four basic results are indicated by the experience to date:

1) One long run is not always the optimal strategy. In many cases, it will be advantageous to consider performing a number of shorter runs rather than a single long run.

2) The optimal number of independent replications increases with increasing traffic intensity.

3) Warm-up is not always worthwhile. In relatively uncongested systems, the cost of discarding data (increased variance) outweighs the benefits (reduced bias) from the standpoint of the mean-square-error criterion. This finding corroborates that of Fishman (4) in his previous work.

4) The greater the traffic intensity, the more useful warm-up is likely to be.

Putting these four results together, we see that it is quite reasonable to consider discarding most of one's available data in the form of warm-up when the system is highly congested, but that in systems with low or medium utilization levels, it is better to use all of one's available data.

AREAS OF CONTINUING RESEARCH

Clearly, the work described in this paper does not represent a complete analysis of the problems of initial condition bias and experimental design. Several areas, in particular, require further extensive study. The first of these is the analysis of sensitivity in the methodology to unknown values of E(X), α, and σ^2. In practice, these will all be unknown quantities for which crude estimates may be available, and a better understanding of the influence of errors in the specification of these values must be obtained.

A second important area is the investigation of initial conditions other than "empty and idle." One can argue, especially for congested systems, that the "empty and idle" start is a very poor choice, and that some alternative choice might lead to much more effective use of experimental resources. This problem is currently under study by the authors, but no conclusive results have been obtained as yet.

Investigation of more complex queuing examples is a third area of high priority for continuing work. The study of simple M/M/1 models has been extremely useful in providing basic insights, but the methods developed must be tested in more complex situations before they can be applied confidently.

Finally, it must be noted that no effort has been made to date to incorporate use of so-called "variance reduction techniques" in the experimental design framework. Such techniques have been studied by a number of authors, and show real promise of effectiveness in the simulation environment. At some point, they should be incorporated into the experimental design framework presented here.

The authors maintain a continuing interest in the problems of experimental design in simulation, and it is hoped that the results reported here will stimulate further research in this important area on the part of other interested simulation analysts.

REFERENCES

1. Eilon, S., and I.G. Chowdhury, "A Note on Steady-State Results in Queuing and Job-Shop Scheduling," Simulation, 23:3 (Sept., 1974), pp. 85-87.

2. Conway, R.W., "Some Tactical Problems in Digital Simulation," Management Science, 10:1 (Oct., 1963), pp. 47-61.

3. Fishman, G.S., "Estimating Sample Size in Computer Simulation Experiments," Management Science, 18:1 (Sept., 1971), pp. 21-38.

4. Fishman, G.S., "A Study of Bias Considerations in Simulation Experiments," to appear in Operations Research.

5. Morse, P.M., Queues, Inventories and Maintenance, New York, Wiley, 1958.