Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

RANKING AND SELECTION FOR STEADY-STATE SIMULATION

David Goldsman

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

William S. Marshall

Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

Seong-Hee Kim
Barry L. Nelson

Department of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.

ABSTRACT

We present and evaluate two ranking-and-selection procedures for use in steady-state simulation experiments when the goal is to find which among a finite number of alternative systems has the largest or smallest long-run average performance. Both procedures extend existing methods for independent and identically normally distributed observations to general stationary output processes, and both procedures are sequential.

1 INTRODUCTION

The "steady-state simulation problem" is one of the central challenges in the design and analysis of stochastic simulation experiments, and it distinguishes simulation experiments from classical statistical experiments. At a high level, the steady-state simulation problem is to estimate some property of a (perhaps vector-valued) random variable that is defined by the limiting distribution of a stochastic process, the limit being taken as the time index of the process goes to infinity. Since the random variable is defined in terms of a limit, realizations of it cannot be obtained (except in special cases that are rarely of practical interest).

In this paper we consider the problem of determining which of a finite number of simulated systems has the largest (or smallest) steady-state mean performance. Our solutions are extensions of existing procedures that have proven performance for the special case of independent and identically distributed (i.i.d.) data >from the normal distribution. As we point out in Section 2, few of the assumptions underlying the existing procedures will be valid in steady-state simulation, particularly when only a single replication is obtained from each system, as we assume. Section 2 also reviews the relevant literature. In Section 3 we describe the new procedures, while Section 4 evaluates them empirically. We conclude by offering our opinions about the key open research questions in Section 5.

2 BACKGROUND

In this section we review the two procedures, designed originally for i.i.d. normal data, that we will extend and enhance for use in steady-state simulation problems. We also more precisely characterize what we mean by the "steady-state simulation problem," and review the literature on ranking and selection (R&S) procedures designed for this case.

2.1 Two Procedures for i.i.d. Normal Data

We describe two procedures that guarantee, with confidence level at least $1-\alpha$, that under certain conditions the system ultimately selected has the largest true mean when the true mean of the best system is at least δ better than the second best. When there are inferior systems whose means are within δ of the true best, then the procedures guarantee to find one of these "good" systems with the same probability. The parameter δ, which defines the *indifference zone*, is set by the experimenter to the smallest absolute difference in expected performance that is considered important to detect. Differences of less than δ are considered practically insignificant. Procedures of this type are known as *indifference-zone ranking and selection procedures*. Comprehensive reviews of ranking and selection procedures can be found in Bechhofer, Santner and Goldsman (1995) and

Goldsman and Nelson (1998). Both procedures studied here—one from Rinott (1978) and the other from Kim and Nelson (2000)—are sequential, by which we mean they typically require two or more stages of simulation.

Suppose that there are $k \geq 2$ systems, and let X_{ij} denote the jth independent observation from system i. Both procedures assume that the $X_{ij} \sim \mathrm{N}(\mu_i, \sigma_i^2)$, with μ_i and σ_i^2 unknown, and that the data across systems are independent. Also let $\bar{X}_i(r) = r^{-1} \sum_{j=1}^{r} X_{ij}$ denote the sample mean of the first r observations from system i.

Rinott's (1978) procedure (RP) requires at most two stages of simulation; it is one of the simplest and most well-known R&S procedures.

The fully sequential procedure (FSP), due to Kim and Nelson (2000), also allows elimination. This procedure takes only a single observation from each alternative system still in play at each stage of simulation, and may choose to cease sampling from systems that no longer appear to be competitive.

Both RP and FSP terminate with a single system that is reported as the best. They could be applied "as is" to steady-state simulation experiments provided we are willing to make multiple replications of each alternative and use the within-replication averages as the basic observations. In the following section we discuss reasons why such an experiment design may not be desirable.

2.2 Steady-State Simulation

Here we define what we mean by "steady-state simulation" and set up the key assumptions.

Now let $X_{i1}, X_{i2}, \ldots$ denote the simulation output process >from the ith alternative system. For example, X_{ij} could be the jth individual waiting time in the ith queueing system under consideration. These observations are typically neither independent—due to the natural dependence in the process—nor identically distributed—due to initializing the process in other than long-run conditions. They are also likely to be non-normal. However, for many processes, appropriate initialization (selection of initial conditions and truncation of some initial data) will yield an output process that approximately satisfies the following collection of assumptions:

Stationarity: $X_{i1}, X_{i2}, \ldots$ forms a stationary stochastic process.

Consistency: $\bar{X}_i(r) \longrightarrow \mu_i$ a.s. as $r \to \infty$.

Functional Central Limit Theorem (FCLT): There exist constants μ_i and $v_i^2 > 0$ such that

$$\frac{\sum_{j=1}^{\lfloor rt \rfloor}(X_{ij} - \mu_i)}{\sqrt{r}} \Longrightarrow v_i W(t)$$

for $0 \leq t \leq 1$, where $W(t)$ is a standard Brownian motion (Weiner) process and $\Longrightarrow$ denotes convergence in distribution as $r \to \infty$.

We will base comparisons on the steady-state means, $\mu_1, \mu_2, \ldots, \mu_k$. Our consistency assumption implies that it is reasonable to estimate μ_i by $\bar{X}_i(r)$ for some suitably large r. What we need to make statistically valid selections in the steady-state simulation environment is a good estimator for the sample mean's variance. This is relatively easy if we make replications, rather than a single long run, but then we have to solve the initialization problem on each replication. This can be very inefficient if large chunks of data need to be deleted from each replication. But worse, if we do a poor job of initializing then we can allow substantial bias to creep into our estimator. By making a single long replication, we minimize the bias.

Rather than directly trying to estimate the $\mathrm{Var}[\bar{X}_i(r)]$, we can instead seek a good estimator of the *variance parameter* (or *asymptotic variance constant*), $v_i^2 \equiv \lim_{r\to\infty} r\mathrm{Var}[\bar{X}_i(r)]$. A number of relevant variance estimation techniques will be discussed in Section 2.3. We incorporate these estimators into extended versions of RP and FSP in Section 3.

2.3 Variance Estimators

In this subsection we will review a few of the popular estimators for the variance parameter v_i^2. These include batch means, overlapping batch means, and various standardized time series estimators. All of the methods rely on the FCLT assumption (and other moment conditions) to produce asymptotically consistent estimates of the variance parameter. All are intuitive: The batch means method uses the sample variance of approximately i.i.d. sample means from contiguous batches of observations. The overlapping batch means technique, as its name suggests, uses a variance estimate based on a weighted sample variance of overlapped batch means — with full knowledge that these means are highly correlated. Standardized time series estimators rely on the fact that properly standardized stationary processes can be readily approximated as Brownian motion.

In all cases, we will work with batches of observations. What will differ among the variance estimators is how the estimation techniques process the batched data.

2.3.1 Batch Means

We can divide n observations, $X_{i1}, X_{i2}, \ldots, X_{in}$, into b contiguous batches, each of length m (where we assume for convenience that $n = bm$); the observations $X_{i,(j-1)m+1}, X_{i,(j-1)m+2}, \ldots, X_{i,jm}$ comprise the jth batch, $j = 1, 2, \ldots, b$. The quantity

$$\bar{X}_{i,j,m} \equiv \frac{1}{m} \sum_{p=1}^{m} X_{i,(j-1)m+p}$$

is called the jth *batch mean* from system i. Under mild conditions, it is known that with $b > 1$ fixed,

$$\begin{aligned} mV_B^2 &\equiv \frac{m}{b-1}\sum_{j=1}^{b}\left(\bar{X}_{i,j,m} - \bar{X}_i(n)\right)^2 \\ &\Longrightarrow \frac{v_i^2\chi^2(b-1)}{b-1}, \end{aligned}$$

as $n \to \infty$ (implying that $m \to \infty$). The symbol $\chi^2(d)$ denotes a chi-squared random variable with d degrees of freedom. It can be shown that if the batch size m and the number of batches b become large in a certain way (Damerdji 1994), then $mV_B^2 \to v_i^2$ almost surely (that is, mV_B^2 is consistent for v_i^2).

2.3.2 Overlapping Batch Means

Instead of working with asymptotically independent batch means as we did above, we now consider *all* batch means of the form

$$\bar{X}_i(j,m) \equiv \frac{1}{m}\sum_{p=0}^{m-1} X_{i,j+p},$$

for $j = 1, 2, \ldots, n-m+1$. The observations $X_{i,j}, X_{i,j+1}, \ldots, X_{i,j+m-1}$ comprise the jth (overlapping) batch from alternative i.

The overlapping batch means (OBM) estimator for the variance parameter v_i^2 is simply

$$mV_O^2 \equiv \frac{nm}{(n-m+1)(n-m)}\sum_{j=1}^{n-m+1}\left(\bar{X}_i(j,m) - \bar{X}_i(n)\right)^2.$$

It can be shown that as the batch size m and the ratio $b \equiv n/m$ become large, the OBM estimator is consistent for v_i^2 (Damerdji 1994). Further, Meketon and Schmeiser (1984) find that the distribution of this estimator is approximated by

$$mV_O^2 \approx \frac{v_i^2\chi^2(d)}{d},$$

where $d = \lfloor 3(b-1)/2 \rfloor$.

2.3.3 Standardized Time Series

We now look at a completely different methodology for estimating v_i^2 known as standardized time series.

For $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, b$, and $h = 1, 2, \ldots, m$, the hth *cumulative mean* from batch j of system i is

$$\bar{X}_{i,j,h} \equiv \frac{1}{h}\sum_{p=1}^{h} X_{i,(j-1)m+p}.$$

For $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, b$, and $0 \le t \le 1$, the *standardized time series* from batch j of system i is given by

$$T_{i,j,m}(t) \equiv \frac{\lfloor mt \rfloor(\bar{X}_{i,j,m} - \bar{X}_{i,j,\lfloor mt \rfloor})}{v_i\sqrt{m}}.$$

Schruben (1983) showed that if $X_{i1}, X_{i2}, \ldots, X_{in}$ is a stationary sequence satisfying certain mild moment and mixing conditions, then as $m \to \infty$ we have $T_{i,j,m}(t) \Rightarrow \mathcal{B}(t)$, $0 \le t \le 1$, a standard Brownian bridge process.

We denote the weighted area under the standardized time series formed by the jth batch of observations >from system i by

$$A_{i,j} \equiv \frac{v_i}{m}\sum_{\ell=1}^{m} w(\ell/m)T_{i,j,m}(\ell/m),$$

where $w(\cdot)$ is a pre-specified weighting function that is continuous on [0,1], not dependent on m, and normalized so that

$$\begin{aligned} &\mathrm{Var}\left(\int_0^1 w(t)\mathcal{B}(t)\,dt\right) \\ &\quad = 2\int_0^1\int_0^u w(u)w(t)t(1-u)\,dt\,du = 1. \end{aligned}$$

This expression can be simplified considerably; see Goldsman, Meketon, and Schruben (1990) for details.

The (weighted) area (A) estimator for v_i^2 is

$$\begin{aligned} mV_A^2 &\equiv \frac{1}{b}\sum_{j=1}^{b} A_{i,j}^2 \\ &\Longrightarrow \frac{v_i^2\chi^2(b)}{b}, \quad b \ge 1. \end{aligned}$$

One may ask: Why bother with the complication of a weighting function? The answer stems from a closer analysis of the small-sample bias of the variance estimators for different choices of the weights. A judicious choice of $w(t)$ can result in the disappearance of the area estimator's first-order bias term, e.g., $w(t) \equiv \sqrt{840}(3t^2 - 3t + 1/2)$, which we use.

2.4 R&S for Steady-State Simulation

The question at hand is how to adapt R&S procedures to steady-state simulation problems. There have been a number of attempts to do so, primarily extending two-stage procedures such as RP. Key to any such extension is a way to characterize the underlying variability of the stochastic output process from each system, typically via an estimator of the asymptotic variance constant v_i^2. Goldsman (1983) and Nakayama (1995) suggest estimating v_i^2 using the batch means method, while Goldsman (1985) proposes methods based on standardized time series. These papers are closest in spirit to our extension of RP.

Iglehart (1977) estimated v_i^2 using the regenerative method, a method that is less generally applicable than the ones we employ. Dudewicz and Zaino (1977) based their estimator of v_i^2 on the assumption that the simulation output process is well represented by an autoregressive order-1 (AR(1)) process, which is clearly not true in general. Sullivan and Wilson (1989) used an estimator of the simulation output spectrum at frequency 0.

Some of these procedures are heuristics, but others have provable asymptotic validity as $\delta \to 0$, which is a strategy that we also employ. Of course, in a real problem δ is a fixed quantity. However, establishing that a procedure is valid in this limiting sense shows that, as we become more and more demanding of the procedure in terms of its ability to distinguish small differences, then we can be more and more confident that the procedure works. This seems like a useful assurance, since selecting the best is most difficult when even tiny differences matter. See also Nakayama (1997) and Damerdji and Nakayama (1996, 1999) for related asymptotic analysis of multiple comparison procedures.

3 NEW PROCEDURES

We now assume that the output from each system, $X_{ij}, i = 1, 2, \ldots, k,\ j = 1, 2, \ldots$, is a stationary stochastic process satisfying the assumptions of Section 2.2, and further that the systems are simulated independently. Implicit in these assumptions are effectively solving any initialization-bias problem, and not using common random numbers to induce dependence across alternatives.

We extend RP and FSP to steady-state simulation by replacing the first-stage variance estimator FSP) with an estimator of the appropriate asymptotic variance constant from Section 2.3. For RP we need an estimator of the marginal asymptotic variance, while for FSP we require an estimator of the asymptotic variance of the difference between pairs of systems.

To be specific, let $X_{i1}, X_{i2}, \ldots, X_{in_0}$ be the first-stage sample from system i. From this sample we form batches of size m and apply a variance estimator V^2 from Section 2.3. In the case of the BM and A estimators we form $b = \lfloor n_0/m \rfloor$ batches of size m; for the OBM estimator we form $n_0 - m + 1$ batches of size m. The degrees of freedom associated with each estimator are $d = b - 1$ for BM, $d = b$ for A, and $d = \lfloor 3(b-1)/2 \rfloor$ for OBM.

Our extension of Rinott's procedure is as follows:

Extended-Rinott's Procedure (ERP)

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \geq 2$ and batch size $m < n_0$.

Initialization: Obtain Rinott's constant $h = h(d, k, 1-\alpha)$ (Bechhofer, Santner, and Goldsman 1995).

Obtain n_0 observations $X_{ij}, j = 1, 2, \ldots, n_0$, from each system $i = 1, 2, \ldots, k$.

For $i = 1, 2, \ldots, k$, compute mV_i^2, the sample asymptotic variance of the data from system i. Let

$$N_i = \max\left\{n_0, \left\lceil \frac{h^2 m V_i^2}{\delta^2} \right\rceil\right\}.$$

Stopping Rule: If $n_0 \geq \max_i N_i$ then stop and select the system with the largest $\bar{X}_i(n_0)$ as the best.

Otherwise, take $N_i - n_0$ additional observations $X_{i,n_0+1}, X_{i,n_0+2}, \ldots, X_{i,N_i}$ from each system i for which $N_i > n_0$.

Select the system with the largest $\bar{X}_i(N_i)$ as the best.

The FSP is modified similarly, as shown below. In the procedure, we estimate the asymptotic variance of the difference, $v_i^2 + v_\ell^2$, by first forming the differenced series $D_{i\ell j} = X_{ij} - X_{\ell j}, j = 1, 2, \ldots$, then applying one of the variance estimators from Section 2.3 to the series $D_{i\ell j}$.

Extended-Fully Sequential Procedure (EFSP)

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \geq 2$ and batch size $m < n_0$. Calculate

$$\eta = \frac{1}{2}\left\{\left[2\left(1 - (1-\alpha)^{1/(k-1)}\right)\right]^{-2/d} - 1\right\}. \qquad (1)$$

Initialization: Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta d$.

Obtain n_0 observations $X_{ij}, j = 1, 2, \ldots, n_0$, from each system $i = 1, 2, \ldots, k$.

For all $i \neq \ell$ compute $mV_{i\ell}^2$, the sample asymptotic variance of the difference between systems i and ℓ. Let

$$N_{i\ell} = \left\lfloor \frac{h^2 m V_{i\ell}^2}{\delta^2} \right\rfloor$$

and let

$$N_i = \max_{\ell \neq i} N_{i\ell}.$$

Here $N_i + 1$ is the maximum number of observations that can be taken from system i.

If $n_0 \geq \max_i N_i + 1$ then stop and select the system with the largest $\bar{X}_i(n_0)$ as the best.

Otherwise set the observation counter $r = n_0$ and go to **Screening**.

Screening: Set $I^{\text{old}} = I$. Let

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \bar{X}_i(r) \geq \bar{X}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{\text{old}}, \ell \neq i \right\}$$

where

$$W_{i\ell}(r) = \max\left\{0, \frac{\delta}{2cr}\left(\frac{h^2 m V_{i\ell}^2}{\delta^2} - r\right)\right\}.$$

Stopping Rule: If $|I| = 1$, then stop and select the system whose index is in I as the best.

Otherwise, take one additional observation $X_{i,r+1}$ from each system $i \in I$ and set $r = r + 1$.

If $r = \max_i N_i + 1$, then stop and select the system whose index is in I and has the largest $\bar{X}_i(r)$ as the best. Otherwise go to **Screening**.

Notice that in EFSP, as in ERP, the variance estimators depend only on the first-stage data. We can show that if $mV^2 \sim v^2\chi^2(d)/d$, then the EFSP achieves the desired probability of correct selection as $\delta \to 0$. However, at best this assumption will be approximately true. Therefore, we also consider a further refinement of the EFSP in which we update the variance estimators as more data are obtained. In order to define the EFSP with updates, we need the concept of a batching sequence m_r: A batching sequence m_r is an integer-valued, nondecreasing function of r with the property that $m_r \longrightarrow \infty$ as $r \longrightarrow \infty$.

Extended-Fully Sequential Procedure with Updates (EFSPU)

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \geq 2$ and initial batch size $m_{n_0} < n_0$. Set $r = n_0$. Calculate

$$\eta = -\log\left[2\left(1 - (1-\alpha)^{1/(k-1)}\right)\right].$$

Initialization: Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta$.

Obtain n_0 observations X_{ij}, $j = 1, 2, \ldots, n_0$, from each system $i = 1, 2, \ldots, k$.

Update: If m_r has changed since the last update, then for all $i \neq \ell$, compute $m_r V_{i\ell}^2(r)$, the sample asymptotic variance of the difference between systems i and ℓ based on b_r batches of size m_r. Let

$$N_{i\ell}(r) = \left\lfloor \frac{h^2 m_r V_{i\ell}^2(r)}{\delta^2} \right\rfloor$$

and let

$$N_i(r) = \max_{\ell \neq i} N_{i\ell}(r).$$

If $r \geq \max_i N_i(r) + 1$ then stop and select the system with the largest $\bar{X}_i(r)$ as the best.

Otherwise go to **Screening**.

Screening: Set $I^{\text{old}} = I$. Let

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \bar{X}_i(r) \geq \bar{X}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{\text{old}}, \ell \neq i \right\}$$

where

$$W_{i\ell}(r) = \max\left\{0, \frac{\delta}{2cr}\left(\frac{h^2 m_r V_{i\ell}^2(r)}{\delta^2} - r\right)\right\}.$$

Stopping Rule: If $|I| = 1$, then stop and select the system whose index is in I as the best.

Otherwise, take one additional observation $X_{i,r+1}$ from each system $i \in I$ and set $r = r + 1$.

If $r = \max_i N_i + 1$, then stop and select the system whose index is in I and has the largest $\bar{X}_i(r)$ as the best. Otherwise go to **Update**.

Under very general conditions we can show that EFSPU is asymptotically valid as $\delta \to 0$.

4 ANALYSIS

In this section we report on a portion of an extensive empirical evaluation of ERP, EFSP and EFSPU. For this study we focus on the ability of a procedure to terminate quickly with a correct selection.

In the study we controlled the number of systems, k; the number of first-stage observations, n_0; the batch size m (or batching sequence m_r); the configuration of the true means, μ_i; and the dependence structure within the process. In all cases system 1 was the true best (had the largest true mean). We obtained the simulation output data from surrogate output processes that allow us to control the

mean and dependence structure of the process, and to initialize the process in steady state. In this paper we report results for the AR(1) process

$$X_{ij} = \mu_i + \phi(X_{i,j-1} - \mu_i) + Z_j$$

where $Z_j \overset{iid}{\sim} \mathrm{N}(0, 1-\phi^2)$ and $-1 < \phi < 1$.

4.1 Configurations and Experiment Design

How strongly the outputs are correlated depends on ϕ. We varied ϕ over the range $-0.3, 0, 0.3, 0.6, 0.9$ to see the performance of the new procedures under various levels of correlation.

When $\phi = 0$, we have independent data and $n_0 = 24$ is an adequate first-stage sample size to obtain variance estimators of good quality. However, we need more data when outputs are highly correlated. To give a fair comparison across different levels of correlation, we chose the first-stage sample size n_0 such that the ratio of the variance of n_0 observations ($v^2(n_0) = n_0 \mathrm{Var}[\bar{X}(n_0)]$) and the asymptotic variance is approximately equal to 1; more specifically, $|1 - v^2(n_0)/v^2| \approx 0.01$. This guarantees that there is enough data so that it is possible to estimate v^2. After n_0 was determined (and it can be determined analytically for the AR(1) process), all divisors of n_0 were employed as batch sizes m, implying n_0/m batches for BM and A, and $n_0 - m + 1$ for OBM.

The number of systems in each experiment varied over $k = 2, 5, 10$.

The indifference-zone parameter was set to $\delta = v_1/\sqrt{n_0}$, where v_1^2 is the asymptotic variance of the best system. Thus, δ is approximately the standard deviation of the first-stage sample mean of the best system.

Two configurations of the true means were used: The slippage configuration (SC), in which μ_1 was set to δ, while $\mu_2 = \mu_3 = \cdots = \mu_k = 0$. To investigate the effectiveness of the procedures in eliminating non-competitive systems, monotone decreasing means (MDM) were also used. In the MDM configuration, the means of all systems were spaced evenly apart δ from the previous mean. In all cases we set the marginal variance of each system to 1.

For each configuration, 1000 macroreplications (complete repetitions) of the entire experiment were performed. In all experiments, the nominal probability of correct selection (PCS) was set at $1 - \alpha = 0.95$.

4.2 Summary of Results

Overall, ERP and EFSP worked well while EFSPU did not.

The nominal PCS, $1 - \alpha$, was attained as long as the batch size was not too small for ERP and EFSP. Tables 1 and 2 show the sample average of the total number of basic observations, and the estimated PCS, of ERP and EFSP, respectively, when SC configurations were tested and the outputs were weakly serially correlated. Tables 3 and 4 show corresponding results when the outputs were highly correlated. Not surprisingly, guaranteeing the nominal PCS required a larger batch size with highly correlated data. For example, the batch size should be larger than 200 when $\phi = 0.9$. On the other hand, batches of size 7 could attain a PCS of at least $1 - \alpha$ when $\phi = 0.3$.

Tables 5 and 6 show corresponding results when MDM configurations were tested and $\phi = 0.9$. Note that PCS under the MDM configuration is higher than under the SC configuration, as expected.

When there is negative serial correlation, our variance estimators tend to overestimate the true variance, causing the PCS of ERP and EFSP to always exceed the nominal level (no results are shown here).

Notice that a choice of large batch size, which helps guarantee achieving the nominal PCS, comes at the cost of a very large total number of observations. Both procedures incorporate the uncertainty of the variance estimator (through their respective h constants) when calculating the sampling requirements, with higher uncertainty increasing the sampling requirements. The large batch-size choice corresponds to a lower-bias, higher-uncertainty estimator of the variance of the data. The higher bias associated with the small batch-size choice accounts for the degraded PCS performance.

When the largest possible batch size was chosen, ERP usually required fewer observations than EFSP did. However, as we took smaller batch size, EFSP outperformed ERP in the sense that the total number of observations used was smaller, without substantially degrading the PCS performance. ERP is not sensitive to the configuration of the means, while EFSP becomes more efficient under the MDM configuration by effectively eliminating inferior systems early in the experimentation (see Tables 4 and 6).

To compare the performance of the different variance estimators, we used BM, OBM and A (with weighting function $w(t)$). BM and OBM have about the same bias as estimators of the asymptotic variance, but for large m and b, OBM's variance is about $1/3$ smaller (Song and Schmeiser 1995). A is first-ordered unbiased, but the variances of BM and A are about the same. The experiments showed that the procedures required fewer observations when OBM and A where employed, instead of BM. The savings were more noticeable when m was large. However, OBM usually had higher PCS and consumed slightly more observations than A, which implied a more conservative procedure. On the other hand, A's PCS deteriorated somewhat more quickly than OBM's as the batch size decreased.

Unfortunately, the performance of EFSPU was not good; in our experiments it did not attain the nominal PCS. The

Table 1: ERP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.3$ and $n_0 = 70$

		SC $k = 2$						SC $k = 5$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
70	1			21121			0.970			168628			0.976
35	2	20734	2796	2874	0.968	0.983	0.958	167861	25176	25977	0.976	1.000	0.970
14	5	1292	1028	1108	0.949	0.948	0.945	7507	5441	6040	0.961	0.965	0.957
10	7	1016	894	921	0.948	0.945	0.943	5363	4454	4695	0.957	0.957	0.952
7	10	868	796	781	0.943	0.940	0.935	4328	3818	3842	0.952	0.948	0.942
5	14	770	729	695	0.939	0.937	0.930	3729	3436	3325	0.948	0.940	0.930
2	35	563	554	1007	0.915	0.913	0.967	2638	2574	4689	0.903	0.901	0.981

Table 2: EFSP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.3$ and $n_0 = 70$

		SC $k = 2$						SC $k = 5$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
70	1			9004			0.977			362921			0.968
35	2	8631	1590	1655	0.969	0.987	0.965	346873	17453	17810	0.968	0.998	0.957
14	5	760	665	480	0.958	0.959	0.940	4762	3766	3751	0.962	0.979	0.955
10	7	601	561	537	0.958	0.955	0.936	3193	2907	2818	0.949	0.968	0.945
7	10	508	487	448	0.948	0.946	0.931	2442	2364	2180	0.941	0.956	0.940
5	14	441	432	388	0.934	0.935	0.927	2068	2019	1814	0.936	0.951	0.930
2	35	304	306	575	0.904	0.910	0.970	1345	1344	2503	0.886	0.892	0.980

Table 3: ERP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.9$ and $n_0 = 1000$

		SC $k = 2$						SC $k = 5$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000	1			71182			0.969			227394			0.974
500	2	70518	38080	37697	0.967	0.982	0.958	227326	197880	182291	0.972	1.000	0.969
250	4	23374	15980	18378	0.949	0.947	0.950	133463	86820	104375	0.961	0.974	0.963
200	5	18379	14875	15874	0.948	0.953	0.948	103709	77466	85710	0.961	0.967	0.960
125	8	13449	12016	12169	0.942	0.939	0.936	69625	59357	60520	0.953	0.949	0.942
100	10	12377	11301	10580	0.947	0.939	0.933	61618	54751	52108	0.952	0.951	0.928
50	20	9742	9409	6331	0.930	0.932	0.886	46466	44112	30072	0.933	0.920	0.838
40	25	9022	8737	5112	0.926	0.918	0.872	42327	40861	24135	0.921	0.923	0.781
25	40	7346	7202	2941	0.911	0.906	0.796	34400	33786	13785	0.885	0.889	0.625
20	50	6558	6472	2291	0.898	0.896	0.769	30566	29969	10227	0.860	0.845	0.541
10	100	4195	4200	2000	0.842	0.844	0.750	19411	19273	5003	0.738	0.730	0.344
8	125	3563	3575	2000	0.822	0.827	0.758	16506	16408	5000	0.680	0.676	0.347
5	200	2448	2435	2000	0.789	0.785	0.759	11276	11212	5000	0.574	0.569	0.344
4	250	2129	2122	2000	0.765	0.763	0.760	9283	9215	5000	0.519	0.518	0.349
2	500	2000	2000	2000	0.763	0.765	0.768	5269	5273	5000	0.381	0.381	0.341

Table 4: EFSP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.9$ and $n_0 = 1000$

		SC $k = 2$						SC $k = 5$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000	1			126041			0.959			5095955			0.960
500	2	119560	23306	24074	0.966	0.975	0.968	4874550	244495	250245	0.969	0.998	0.967
250	4	13831	10994	10713	0.956	0.952	0.944	103105	67388	69124	0.964	0.976	0.962
200	5	10915	9588	9475	0.940	0.942	0.942	67271	53183	53352	0.954	0.977	0.965
125	8	7946	7536	7166	0.940	0.939	0.934	40957	37768	34832	0.950	0.968	0.932
100	10	7139	6935	6065	0.931	0.940	0.924	34962	33562	29251	0.942	0.957	0.925
50	20	5437	5424	3558	0.919	0.920	0.869	24843	24554	15173	0.921	0.929	0.818
40	25	4976	4977	2957	0.910	0.916	0.828	22362	22230	11659	0.910	0.913	0.758
25	40	3996	3995	2169	0.885	0.884	0.782	17298	17233	6806	0.892	0.870	0.646
20	50	3615	3613	2054	0.872	0.871	0.766	15061	15051	5702	0.836	0.846	0.579
10	100	2476	2472	2000	0.815	0.809	0.749	8954	8946	5002	0.703	0.707	0.497
8	120	2275	2276	2000	0.788	0.790	0.749	7635	7635	5000	0.670	0.668	0.496
5	200	2064	2063	2000	0.765	0.765	0.749	5847	5850	5000	0.586	0.590	0.496
4	250	2028	2028	2000	0.756	0.755	0.749	5440	5439	5000	0.555	0.554	0.496
2	500	2000	2000	2000	0.749	0.749	0.749	5022	5022	5000	0.499	0.498	0.496

Table 5: ERP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.9$ and $n_0 = 1000$

		MDM $k = 2$						MDM $k = 10$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000	1			227439			0.990			476269			0.994
500	2	227293	198000	182388	0.989	1.000	0.987	476523	465164	423794	0.995	1.000	0.994
250	4	133365	86683	104241	0.985	0.989	0.988	345137	239476	285257	0.992	0.995	0.994
200	5	103616	77483	85654	0.986	0.987	0.986	284246	211784	239353	0.992	0.996	0.993
125	8	69617	59302	60542	0.986	0.981	0.982	190422	157110	163794	0.990	0.992	0.991
100	10	61562	54751	52103	0.984	0.982	0.976	165393	144198	139748	0.993	0.991	0.988
50	20	46472	44112	30080	0.977	0.976	0.953	121198	114408	78225	0.988	0.986	0.967
40	25	42340	40828	24146	0.976	0.977	0.932	110617	105759	62608	0.988	0.986	0.950
25	40	34384	33786	13783	0.960	0.964	0.865	88956	86658	35666	0.980	0.980	0.906
20	50	30573	29969	10227	0.956	0.956	0.834	79192	77045	26296	0.974	0.970	0.868
10	100	19407	19281	5003	0.915	0.911	0.738	49901	49622	10307	0.942	0.935	0.728
8	125	16507	16401	5000	0.891	0.893	0.735	42459	42178	10000	0.921	0.920	0.724
5	200	11273	11213	5000	0.843	0.846	0.725	28828	28694	10000	0.882	0.878	0.729
4	250	9282	9214	5000	0.817	0.818	0.734	23736	23751	10000	0.850	0.853	0.732
2	500	5270	5273	5000	0.731	0.735	0.722	12765	12759	10000	0.762	0.768	0.736

Table 6: EFSP's Sample Average Total Number of Basic (Unbatched) Observations and PCS When AR(1) Processes Are Tested With $\phi = 0.9$ and $n_0 = 1000$

		MDM $k = 2$						MDM $k = 10$					
		Observations			PCS			Observations			PCS		
m	b	BM	OBM	A	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000	1			3094055			0.986			16678305			0.993
500	2	3006572	150354	154090	0.987	1.000	0.985	16863872	394457	384010	0.994	1.000	0.997
250	4	61829	42802	42919	0.985	0.992	0.988	128588	80301	82144	0.990	0.997	0.995
200	5	41548	33898	33358	0.986	0.991	0.988	78519	60704	60317	0.996	0.998	0.993
125	8	25892	23887	22127	0.982	0.987	0.983	44429	40656	38477	0.991	0.995	0.989
100	10	22241	21245	18315	0.978	0.984	0.982	37461	35593	31254	0.995	0.994	0.989
50	20	15727	15773	10534	0.975	0.979	0.943	26254	26087	18296	0.988	0.989	0.962
40	25	14476	14372	8569	0.972	0.975	0.912	23930	23759	15438	0.982	0.984	0.946
25	40	11526	11535	5999	0.958	0.952	0.832	19496	19416	11607	0.974	0.975	0.867
20	50	10286	10268	5380	0.943	0.943	0.785	17666	17636	10710	0.964	0.967	0.809
10	100	7125	7127	5001	0.891	0.891	0.721	13161	13171	10006	0.921	0.918	0.735
8	120	6398	6404	5000	0.863	0.862	0.718	12226	12210	10000	0.899	0.897	0.733
5	200	5486	5485	5000	0.788	0.787	0.718	10873	10872	10000	0.824	0.825	0.732
4	250	5267	5267	5000	0.771	0.773	0.718	10476	10476	10000	0.802	0.802	0.732
2	500	5015	5015	5000	0.725	0.725	0.718	10042	10042	10000	0.744	0.743	0.732

key factors in the performance of our fully sequential procedures are the variance estimators and the parameter η. In EFSPU, the computation of η is based on the assumption that the variance estimator is very close to the true asymptotic variance; essentially, we are pretending that v^2 is known. Although this is asymptotically valid as $\delta \to 0$, in practice it results in a continuation region that is too narrow, leading to a greater than desired chance that good systems are eliminated. For instance, even with independent data the PCS of EFSPU was around 0.8. We are investigating modifications of this procedure that use a larger value of η because we believe that variance updating has the potential to lead to improved procedures.

5 THE FUTURE

The empirical evidence presented here, as well as other analysis we have undertaken, convinces us that R&S procedures can be applied to steady-state simulation problems in which only a single replication is obtained from each system. Procedure ERP has the advantage that data can be collected from each system without reference to the others, making it easy to implement in distributed computing environments. EFSP and EFSPU are highly efficient procedures, but they assume the ability to obtain incremental output data from each system as needed.

Despite our confidence, there are a number of issues yet to be resolved:

- The longstanding initialization-bias problem is at least as critical here as it is in estimating parameters of a single system.
- Even assuming the initialization-bias problem is solved, there is still a fundamental question of when enough data have been collected to have a statistically valid first-stage sample (what we call n_0). For ERP and EFSP, enough data must be collected to have an approximately (scaled) chi-squared variance estimator with low bias. When data are highly dependent this is difficult to determine. Since EFSPU updates the variance estimators, it may be able to overcome errors in determining an acceptable initial sample size or batch size provided it does not terminate too early.
- None of the new procedures introduced here directly incorporate the variance reduction technique of common random numbers (CRN). CRN can be effective at reducing the sample size required to reach a correct selection, as shown in Kim and Nelson (2000) for FSP. Because CRN induces dependence across systems, and we already have dependence within replications, it becomes difficult to provide procedures that account for both.

REFERENCES

Bechhofer, R. E., T. J. Santner and D. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. New York: John Wiley.

Damerdji, H. 1994. Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research* 19:494–512.

Damerdji, H. and M. K. Nakayama. 1996. Two-stage procedures for multiple comparisons with a control in steady-state simulations. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 372–375. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Damerdji, H. and M. K. Nakayama. 1999. Two-stage multiple-comparison procedures for steady-state simulations. *ACM TOMACS* 9:1–30.

Dudewicz, E. J. and N. A. Zaino. 1977. Allowance for correlation in setting simulation run-length via ranking-and-selection procedures. *TIMS Studies in the Management Sciences* 7:51–61.

Goldsman, D. 1983. Ranking and selection in simulation. In *Proceedings of the 1983 Winter Simulation Conference*, ed. S. Roberts, J. Banks, and B. Schmeiser, 387–393. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Goldsman, D. 1985. Ranking and selection procedures using standardized time series. In *Proceedings of the 1985 Winter Simulation Conference*, ed. D. Gantz, G. Blais, and S. Solomon, 120–124. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Goldsman, D., M. S. Meketon, and L. W. Schruben. 1990. Properties of standardized time series weighted area variance estimators. *Management Science* 36:602–612.

Goldsman, D. and B. L. Nelson. 1998. Comparing systems via simulation. In *Handbook of Simulation*, ed. J. Banks, Chapter 8. New York: John Wiley.

Iglehart, D. L. 1977. Simulating stable stochastic systems, VII: Selecting the best system. In *TIMS Studies in the Management Sciences* 7:37–49.

Kim, S-H. and B. L. Nelson. 2000. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS*, forthcoming.

Meketon, M. S. and B. Schmeiser. 1984. Overlapping batch means: Something for nothing? In *Proceedings of the 1984 Winter Simulation Conference*, ed. S. Sheppard, U. Pooch and D. Pedgen, 227–230. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Nakayama, M. K. 1995. Selecting the best system in steady-state simulations using batch means. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 362–366. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Nakayama, M. K. 1997. Multiple-comparison procedure for steady-state simulations. *Annals of Statistics* 25:2433–2450.

Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Comm. Stat.—Thy. and Meth.* A7:799–811

Schruben, L. W. 1983. Confidence interval estimation using standardized time series. *Operations Research* 31:1090–1108.

Song, W. T. and B. W. Schmeiser. 1995. Optimal mean-squared-error batch sizes. *Management Science* 41:110–123.

Sullivan, D. W. and J. R. Wilson. 1989. Restricted subset selection procedures for simulation. *Operations Research* 37:52–71.

AUTHOR BIOGRAPHIES

DAVID GOLDSMAN is a Professor in the School of Industrial and Systems Engineering at Georgia Tech. His research interests include simulation output analysis and ranking and selection. He is the Simulation Department Editor for *IIE Transactions*, and an Associate Editor for *Operations Research Letters*. He was the Program Chair for the 1995 Winter Simulation Conference. His e-mail and web addresses are <sman@isye.gatech.edu> and <www.isye.gatech.edu/~sman/>.

SEONG-HEE KIM is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. She received her B.S. in industrial management from KAIST. Her research interests include ranking and selection procedures and simulation output analysis of stochastic systems. Her e-mail address is <shkim@iems.nwu.edu>.

WILLIAM S. MARSHALL is a Research Engineer at the Georgia Tech Research Institute, having recently obtained his Ph.D. from the School of Industrial and Systems Engineering at Georgia Tech. His research interests include radar signal analysis and tracking algorithms, radar system simulation, and ranking and selection procedures. His e-mail address is <bill.marshall@gtri.gatech.edu>.

BARRY L. NELSON is a Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University, and is Director of the Master of Engineering Management Program there. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. He has published numerous papers and two books. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and current member of the Board of Directors. His e-mail and web addresses are <nelsonb@northwestern.edu> and <www.iems.northwestern.edu/~nelsonb/>.