Research Perspective

Ranking and Selection for Steady-State Simulation: Procedures and Perspectives

David Goldsman • Seong-Hee Kim • William S. Marshall • Barry L. Nelson
School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, Georgia 30332
Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois 60208-3119
sman@isye.gatech.edu • skim@isye.gatech.edu • bill.marshall@gtri.gatech.edu • nelsonb@northwestern.edu

We present and evaluate three 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. All three procedures extend existing methods for independent and identically normally distributed observations to general stationary output processes, and all procedures are sequential. We also provide our thoughts about the evaluation of simulation design and analysis procedures, and illustrate these concepts in our evaluation of the new procedures.

(*Simulation, Design of Experiments; Simulation, Statistical Analysis; Statistics, Time Series*)

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 operations research, management science and industrial engineering contexts, steady-state simulation problems arise in the design of manufacturing, service and information systems when the planning horizons are long or time-dependent behavior is not relevant.

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 in which the observations from each system are independent and identically distributed (i.i.d.) data from a 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. Section 3 contains

0899-1499/02/1401/0002$5.00
1526-5528 electronic ISSN

our perspective on how a researcher can establish that a new statistical procedure is useful. In Section 4 we describe the new procedures, while in Sections 5 and 6 we evaluate them based on the ideas described in Section 3. We conclude by offering our opinions about key open research questions in Section 7.

2. Background

In this section we review 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 define 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 "close enough" 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 R&S procedures. Comprehensive reviews of R&S can be found in Bechhofer et al. (1995), to which we henceforth refer as BSG 1995, and Goldsman and Nelson (1998). Both procedures studied here—one from Rinott (1978) and the other from Kim and Nelson (2001a)—are sequential, by which we mean they typically require two or more stages of simulation. When process variances are unknown—which is almost always the case in real problems—then at least two stages of sampling are required to deliver a prespecified probability of correct selection.

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 $X_{ij} \sim N(\mu_i, \sigma_i^2)$, 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 requires at most two stages of simulation; it is one of the simplest and most well-known R&S procedures.

Rinott's Procedure ($\mathcal{R}$)

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

Initialization: Obtain Rinott's constant $h = h(n_0, k, 1-\alpha)$ (for instance, from BSG 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

$$S_i^2 = \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}(X_{ij} - \bar{X}_i(n_0))^2,$$

the sample variance of the data from system i. Let

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

where $\lceil\cdot\rceil$ indicates rounding up any fractional part to the next larger integer. Here N_i is the number of observations that will be taken from system i.

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 following fully sequential procedure is due to Kim and Nelson (2001a). This procedure takes only a single observation from the systems still in play at each stage of simulation, and may choose to cease sampling from systems that no longer appear to be competitive.

Kim and Nelson's Procedure ($\mathcal{KN}$)

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

$$\eta = \frac{1}{2}\{[2(1-(1-\alpha)^{1/(k-1)})]^{-2/(n_0-1)} - 1\}.$$

Initialization: Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta(n_0 - 1)$. 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

$$S_{i\ell}^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_{ij} - X_{\ell j} - [\overline{X}_i(n_0) - \overline{X}_\ell(n_0)])^2,$$

the sample variance of the difference between systems i and ℓ. Let

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

where $\lfloor \cdot \rfloor$ indicates truncation of any fractional part, 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 $\overline{X}_i(n_0)$ as the best.

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

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

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

where

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

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 $\overline{X}_i(r)$ as the best. Otherwise go to **Screening**.

Both $\mathscr{R}$ and $\mathscr{KN}$ 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 a single replication of 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; see, for instance, Law and Kelton 2000) will yield an output process that approximately satisfies the following collection of assumptions:

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

(Strong) Consistency: $\overline{X}_i(r) \longrightarrow \mu_i$ a.s. (almost surely) as $r \to \infty$.

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

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

for $0 \leq t \leq 1$, where $\mathscr{W}(t)$ 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 $\overline{X}_i(r)$ 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 mitigate the bias problem.

Rather than directly estimating the var$[\overline{X}_i(r)]$, we can instead seek a good estimator of the *variance parameter* (or *asymptotic variance constant*), $v_i^2 \equiv$

$\lim_{r\to\infty} r\,\mathrm{var}[\overline{X}_i(r)]$. A number of relevant variance estimation techniques for doing this will be discussed in Section 2.3. We incorporate these estimators into extended versions of $\mathscr{R}$ and $\mathscr{KN}$ in Section 4.

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 estimators of the variance parameter. 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

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

is called the jth *batch mean* from system i. Under mild conditions (e.g., Glynn and Whitt 1991, Steiger and Wilson 2001), it is known that with $b > 1$ fixed,

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

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. We refer to mV_B^2 as the batch means (BM) estimator. It can be shown that if the batch size m and the number of batches b both become large in a certain way (Damerdji 1994), then $mV_B^2 \to v_i^2$ almost surely (that is, mV_B^2 is strongly consistent for v_i^2; see Chien et al. 1997 for complementary mean-square consistency results).

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

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

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n-m+1$. 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

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

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 well 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

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

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

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

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 $T_{i,j,m}(t) \Longrightarrow \mathscr{B}(t), 0 \le t \le 1$, 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

$$\operatorname{var}\left(\int_0^1 w(t)\mathcal{B}(t)\,dt\right) = 2\int_0^1\int_0^u w(u)w(t)t(1-u)\,dt\,du = 1.$$

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

$$mV_A^2 \equiv \frac{1}{b}\sum_{j=1}^{b} A_{i,j}^2 \Longrightarrow \frac{v_i^2\chi^2(b)}{b}$$

for fixed $b \geq 1$ as $m \to \infty$.

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. It can be shown (Goldsman et al. 1990 and Song and Schmeiser 1995) that, as estimators of v_i^2,

$$\begin{aligned}\text{Bias}(mV_B^2) &= \frac{\gamma_i}{m} + o(1/m),\\ \text{Bias}(mV_O^2) &= \frac{\gamma_i}{m} + o(1/m), \text{ and}\\ \text{Bias}(mV_A^2) &= \frac{f(w)\gamma_i}{m} + o(1/m),\end{aligned}$$

where γ_i is a constant that depends only on the autocorrelation structure of the stochastic process underlying the ith system; $o(1/m)$ indicates convergence to zero more quickly than $1/m$ as the batch size m becomes large; and $f(w)$ is a function of the area estimator's weighting. 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 in this paper.

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 $\mathcal{R}$. 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 $\mathcal{R}$.

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. However, our limiting argument, which is applied to $\mathcal{KN}$, is fundamentally different from their argument.

3. Perspectives on Evaluating Procedures

When new procedures for the design and analysis of simulation experiments are proposed, the inventor has a fundamental obligation to establish the "goodness" of the procedures. There are many techniques for doing this, and in most cases multiple approaches are required. In this section we provide our perspectives on this important task, and illustrate them by our evaluation of the R&S procedures proposed in this paper.

Suppose that there is a new procedure, which we denote by $\mathcal{P}$, that is under review. We vaguely define "procedure" to be anything from an experiment design strategy to an output analysis method, or a combination of the two. We assume only that $\mathcal{P}$, when applied to a simulation, is supposed to provide some information about the simulation model.

Why is it necessary to evaluate $\mathcal{P}$ separately from its actual use? Typically, $\mathcal{P}$ is proposed because the researcher has some mathematical, empirical or intuitive justification for believing that it is useful. If we can show mathematically—by formal statement of conditions and proof of the claims—that $\mathcal{P}$ performs as desired under all conditions for which it will be employed, then the evaluation is complete and the user can apply $\mathcal{P}$ with confidence. We refer to this as an *exhaustive mathematical analysis* because it must cover all conditions of interest.

More often, the best that we can do is show that $\mathcal{P}$ has *some* desirable properties under conditions that are more restrictive than we can be certain we will encounter in practice. For example, the new R&S procedures in this paper work as advertised if the output data from each of the simulated systems are i.i.d. normal. Obviously this condition is never precisely true in real applications; in fact, we hope that these procedures can be used when the i.i.d. normal condition is violated. Here "work as advertised" means that the procedures select the system with the largest true mean, or one that is close to the best (in a precisely defined way), with a prespecified probability.

Further complicating the matter, the "goodness" of $\mathcal{P}$ is rarely captured by a single performance measure, and we may be interested in $\mathcal{P}$'s performance relative to other procedures, not just in an absolute sense. A R&S procedure, for instance, is supposed to provide an absolute guarantee of achieving the desired probability of correct selection (PCS). If, however, there are other procedures that may be applied to the same problem, then the value of $\mathcal{P}$ could be that it is more efficient in terms of its expected sample size than the competing procedures. In simulation methodology research it is rare that the performance of a procedure can easily be evaluated on all dimensions of interest, and also rare that a new procedure dominates all existing procedures on every relevant performance measure.

When an exhaustive mathematical analysis is not possible, then there are at least four other forms of analysis that are appropriate: *mathematical analysis under idealized conditions; analysis under surrogate models; asymptotic analysis*; and *empirical analysis*. We do not claim that these categories include all possible approaches, nor that they are mutually exclusive. But we do believe that they cover the key techniques. In the remainder of this section we comment on each one.

We can sometimes perform the analysis we want, restricted to *idealized conditions*, when exhaustive mathematical analysis is not possible. This might also be called "evaluating special cases," but more broadly it means that we examine the procedure exhaustively over either a limited list of performance measures or under restrictive assumptions.

For instance, in Kim and Nelson (2001a) we studied the impact of batching on $\mathcal{KN}$. The idealized condition we assumed was that a sufficient condition to guarantee the desired PCS—the batch means being i.i.d. normal—holds at all batch sizes. By doing this we could perform an exhaustive analysis of the impact of batch size on the efficiency of the procedure (number of observations until termination). Specifically, we determined whether or not it is important to find the smallest possible batch size that works in the first-stage sample. The answer is that it is not critical as long as a certain minimal number of batches are obtained. The result is useful, even though based on idealized conditions, because the fact that it is not necessary to obtain the smallest allowable batch size when conditions are most favorable for that strategy suggests that it is not essential when the conditions are unfavorable.

In simulation research there is a long history of employing relatively simple models as surrogates for the simulation model itself; for this reason we separate *surrogate models* from analysis under idealized conditions. These models, while simple, are chosen because they share some important characteristics with the real simulations they represent. In simulation output analysis the M/M/1 queue and autoregressive moving-average (ARMA) time-series models are often used for this purpose. The primary reasons for using surrogate models are that they share key characteristics with real simulations (e.g., statistically dependent output data) while also allowing some control of these characteristics (e.g., different settings for an ARMA model's parameters lead to different levels of dependence).

Asymptotic analysis typically means analysis as the simulation effort (run length, number of replications, or perhaps both) increases (conceptually) without bound. The power of asymptotic analysis is that many of the problem-specific details that thwart mathematical analysis in the finite-sample case wash out in the limit. Asymptotic analysis, done appropriately, can cleanly establish the large-sample validity of a procedure, or the asymptotic superiority of one procedure over another. Further, asymptotic analysis can sometimes lead to modifications that improve small-sample performance.

There are at least two tricky aspects of asymptotic analysis, one technical and one practical. First, there are a number of ways in which the simulation effort can "get large," and a number of ways to look at what happens when it does. For instance, the variance of any sensible point estimator will go to zero as the sample size goes to infinity, but that does not mean that all point estimators are equally good. Scaling up the variance at the same rate at which it is going to zero can sometimes reveal important differences among estimators.

In Kim and Nelson (2001b) we establish the asymptotic validity of our new fully sequential procedures. In Section 5 of this paper, we will use asymptotic analysis, in conjunction with surrogate models, to assess the impact of bias in our variance estimators on the performance of the fully sequential procedure. We drive the run length to infinity by letting the indifference zone δ and the true differences among the systems' means go to 0. Therefore, we can interpret the asymptotic results as telling us what will happen as the problem becomes more and more difficult, which is what we would like to know (few errors occur in easy problems where the means are widely different).

The second tricky point is determining what the large-sample performance tells us about how a procedure will perform when it is actually applied in less-than-infinite samples. In other words, how large is "large enough" for asymptotic performance to be representative of actual performance? Sometimes convergence rates can be determined, but even then there may be unknown constants involved that prevent us from saying much about a specific sample size. Thus, some other form of analysis is needed to support the asymptotic results, and this is often empirical.

Empirical analysis is perhaps the most general technique at our disposal. We make a distinction between empirical analysis and illustrative examples. Illustrative examples play an important role because they demonstrate how a procedure is implemented and how the results might be interpreted; what they lack is control of the factors that might affect procedure performance, and control of the error in the evaluation itself. Control is a key feature of empirical analysis, because the goal of empirical analysis is to make statements about cases we do not examine based on cases that we do. Without control, empirical results can rarely be generalized beyond the specific instances that were evaluated.

Controlled experimentation is a topic that is well known in statistical experiment design. The idea is to identify the factors that might affect the performance of $\mathcal{P}$, both favorably and unfavorably, and vary them in a systematic fashion. The same is true, for instance, in industrial experiments. However, in industrial experiments the range for each experimental factor, such as temperature, pressure, speed, etc., is often determined by the physical nature of the process. This is not necessarily true when evaluating a design and analysis procedure that is supposed to apply to any problem within a rather large and diverse class. The solution to this dilemma is to link factors in meaningful ways.

For instance, when evaluating R&S procedures, factors such as the indifference-zone size, spacing of the true means and standard deviation of the output data affect procedure efficiency. Rather than set the levels of these three factors independently, we can fix the standard deviation of the data from the best system, measure the width of the indifference zone in units of this standard deviation, and measure the spacings of the means in units of the indifference zone. This allows us to draw general conclusions about how the procedures perform when the indifference zone or spacings are large or small relative to the variability of the data, and to avoid tying results to specific values of any of these.

Control over the key factors often comes at a price: models that allow for a great deal of control are usually idealized and may (unknown to us) lack some features of real problems that have a profound impact on performance. For instance, in the empirical evaluation in Section 6 of this paper we use ARMA models as surrogates for the simulation output data. This has the advantage of giving us control over the means, variances and strength of autocorrelation in the simulation output processes. And while we can argue that these time-series models share some characteristics with real problems, we cannot be certain that they include all of the characteristics we might find, say, in the cycle times generated by a complex manufacturing simulation. For this reason, including a few models that are perhaps less controllable, but a step closer to realism, is often appropriate.

Just controlling the key factors is not enough, however, when empirical evaluation of performance is estimated through repeated trials. For instance, we will evaluate the ability of our R&S procedures to provide a prespecified PCS by applying them repeatedly to a situation in which the identity of the best system is known to us. Our estimate of the true PCS is simply the number of times the procedure is correct divided by the number of trials. The number of trials we perform then determines how precise our estimate of PCS is, and it should be chosen to insure adequate precision.

Selecting the number of trials or "macroreplications" is relatively easy when estimating PCS because an unbiased estimator of the standard error of a proportion is readily available. Unfortunately, this is not always the case. When the error associated with a performance estimate is not easy to evaluate, we can add another layer of macroreplications whose purpose is to quantify the error in the performance estimate, rather than the performance of $\mathcal{P}$ itself. We refer to this layering as an "experiment within an experiment."

To make the concept more concrete, let $\mathcal{I}$ denote a random problem instance to which $\mathcal{P}$ might be applied. Further, let $\mathcal{M}_1$ denote a performance measure associated with $\mathcal{P}$; examples include a mean value, bias, standard error, PCS, etc. Our goal is to determine $\mathcal{M}_1(\mathcal{P}(\mathcal{I}))$, a performance attribute of $\mathcal{P}$ when applied to instances of type $\mathcal{I}$. However, we also want to make sure that our estimate of $\mathcal{M}_1$ has satisfactory performance, as determined by measure $\mathcal{M}_2$ (standard error, for instance). This might lead to the following set of nested experiments:

```
    $\vdots$
for $\ell$ from 1 to $r_3$ do
    for $j$ from 1 to $r_2$ do
        for $i$ from 1 to $r_1$ do
            generate instance $\mathcal{I}$
            apply $\mathcal{P}$ to $\mathcal{I}$
        loop
        $\widehat{\mathcal{M}}_1$ estimates $\mathcal{M}_1(\mathcal{P}(\mathcal{I}))$
    loop
    $\widehat{\mathcal{M}}_2$ estimates $\mathcal{M}_2(\widehat{\mathcal{M}}_1)$
loop
$\widehat{\mathcal{M}}_3$ estimates $\mathcal{M}_3(\widehat{\mathcal{M}}_2)$
    $\vdots$
```

The $\vdots$ indicates that, if desired, we could add even more layers to the experiment, with each new layer allowing us to evaluate the performance estimate from the inner stage. Typically two to three layers are adequate, but the fact that we can use an arbitrary number of layers is a distinct advantage of empirical evaluation in simulation.

4. 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 the technique of common random numbers to induce dependence across alternatives.

We extend $\mathcal{R}$ and $\mathcal{KN}$ to steady-state simulation by replacing the first-stage variance estimator (S_i^2 for $\mathcal{R}$, $S_{i\ell}^2$ for $\mathcal{KN}$) with an estimator of the appropriate asymptotic variance constant from Section 2.3. For $\mathcal{R}$

we need an estimator of the marginal asymptotic variance, while for $\mathcal{KN}$ 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 mV^2 from Section 2.3. In the case of the BM and A estimators we form $b_0 = \lfloor n_0/m_0 \rfloor$ batches of size m_0; for the OBM estimator we form $n_0 - m_0 + 1$ batches of size m_0. The degrees of freedom associated with each estimator are $d = b_0 - 1$ for BM, $d = b_0$ for A, and $d = \lfloor 3(b_0 - 1)/2 \rfloor$ for OBM.

4.1. Extended Rinott's Procedure ($\mathcal{R}+$)

Our extension to Rinott's procedure is as follows:

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \geq 2$ and batch size $m_0 < n_0$.

Initialization: Obtain Rinott's constant $h = h(d, k, 1-\alpha)$ (BSG 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 $m_0 V_i^2$, the sample asymptotic variance of the data from system i. Let

$$N_i = \max\left\{n_0, \left\lceil \frac{h^2 m_0 V_i^2}{\delta^2} \right\rceil\right\}.$$

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.

4.2. Extended Kim and Nelson's Procedure ($\mathcal{KN}+$)

$\mathcal{KN}$ is modified analogously to $\mathcal{R}$. 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 $D_{i\ell j}$.

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \geq 2$ and batch size $m_0 < n_0$. Calculate

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

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 $m_0 V_{i\ell}^2$, the sample asymptotic variance of the difference between systems i and ℓ. Let

$$N_{i\ell} = \left\lfloor \frac{h^2 m_0 V_{i\ell}^2}{\delta^2} \right\rfloor$$

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

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

where

$$W_{i\ell}(r) = \max\left\{0, \frac{\delta}{2r}\left(\frac{h^2 m_0 V_{i\ell}^2}{\delta^2} - r\right)\right\}.$$

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 $\mathcal{KN}+$, as in $\mathcal{R}+$, the variance estimators depend only on the first-stage data. In Kim and Nelson (2001b) we show that if $m_0 V_{i\ell}^2 \sim v_{i\ell}^2 \chi^2(d)/d$, then $\mathcal{KN}+$ achieves the desired probability of correct selection as $\delta \to 0$. However, this distribution assumption will be approximately true at best, and so we also consider a more dramatic refinement of $\mathcal{KN}$ in which we *update the variance estimators as more data are obtained.*

4.3. Extended Kim and Nelson's Procedure with Updates ($\mathcal{KN}++$)

To define this new procedure, we first need the concept of a *batching sequence*: A batching sequence is defined as $\{(m_r, b_r)\}$ where m_r is an integer-valued, nondecreasing function of the sampling stage r with the property that $m_r \le r/2$, and $m_r \longrightarrow \infty$ as $r \longrightarrow \infty$. The function m_r is the batch size when r observations have been created; the number of batches is $b_r = \lfloor r/m_r \rfloor$ for batch means and area estimators, and $b_r = r - m_r + 1$ for overlapping batch means. Specific examples of batching sequences are provided in Section 6.

Setup: Select confidence level $1-\alpha$, indifference-zone parameter $\delta > 0$, first-stage sample size $n_0 \ge 2$ and initial batch size $m_0 < n_0$. Calculate

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

Initialization: Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta d$. (Note that since d is a function of the number of batches, b_r, the value of h^2 will change whenever b_r changes.)

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

Set the observation counter $r = n_0$ and $m_r = m_0$.

Update: If m_r has changed since the last update ($m_r \ne m_{r-1}$), then for all $i \ne \ell$, 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

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

If $r \ge \max_i N_i(r) + 1$ 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

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

where

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

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 $r = \max_i N_i(r) + 1$, 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 Kim and Nelson (2001b) show that $\mathcal{KN}++$ is asymptotically valid as $\delta \to 0$.

5. Asymptotic Analysis and Surrogate Models

The asymptotic validity of $\mathcal{KN}+$, as shown in Kim and Nelson (2001b), is based on the assumption that the variance estimators computed from the first stage sample follow $v_{i\ell}^2\chi^2(d)/d$ distributions, where $v_{i\ell}^2 = v_i^2 + v_\ell^2$ is the asymptotic variance of the difference between systems i and ℓ, and d depends on the variance estimator (we refer to this as Assumption C). However, this is never precisely true, and violation of the assumption affects the validity of the procedure. In this section, we analyze how asymptotic performance is affected by one type of deviation from this assumption.

Let b_0 and m_0 denote the number of batches and batch size, respectively, and $n_0 = m_0 b_0$ the first-stage sample size. Where there will be no confusion, we drop the subscript 0 in this section so that m and b also refer to first-stage batch size and number of batches, respectively. Let $\bar{X}_{i,1,m}, \bar{X}_{i,2,m}, \ldots, \bar{X}_{i,b,m}$ be the first-stage batch means of size m from system i. We define $\text{var}(\bar{X}_{i,j,m})$ to be the variance of a batch mean of size m from system i, and define $v_i^2(m)$ to be

$$v_i^2(m) \equiv m\text{var}(\bar{X}_{i,j,m}).$$

Thus, $v_i^2 = \lim_{m\to\infty} v_i^2(m)$, but $v_i^2 \ne v_i^2(m)$ in general for finite m. The effect of this bias is what we will examine.

If m is large enough, then the $\{\bar{X}_{i,j,m}, j = 1, 2, \ldots, b\}$ are approximately i.i.d. $\text{N}(\mu_i, v_i^2(m)/m)$, so that

$$mV_B^2 \approx \frac{v_i^2(m)\chi^2(b-1)}{b-1} \tag{1}$$

where $\approx$ means "approximately distributed as." Further, from the discussion in Section 2.3.2, we have

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

where $d = \lfloor 3(b-1)/2 \rfloor$. For the area estimator, if the $A_{i,j}^2$'s are assumed to be approximately i.i.d. $v_i^2(m)\chi^2(1)$, then

$$mV_A^2 \approx \frac{v_i^2(m)\chi^2(b)}{b}. \tag{3}$$

In Kim and Nelson (2001b) we showed that, under Assumption C, the asymptotic probability of an incorrect selection (ICS) when $k \geq 2$, $\mu_k - \delta = \mu_{k-1} = \cdots = \mu_1$, and $\delta \to 0$, is

$$\Pr\{\text{ICS}\} \longrightarrow 1 - \left[1 - \frac{1}{2}(1+2\eta)^{-d/2}\right]^{k-1} = \alpha$$

where the final equality is a result of the way we select η. However, if we only assume that the batch size is large enough that (1), (2) or (3) holds instead of Assumption C, then the asymptotic probability of incorrect selection (APICS), as $\delta \to 0$, is

$$\text{APICS} = 1 - \prod_{i=1}^{k-1}\left[1 - \frac{1}{2}\left(1 + 2\eta\frac{v_{ik}^2(m)}{v_{ik}^2}\right)^{-d/2}\right]$$

where $v_{ik}^2 = v_i^2 + v_k^2$ is the asymptotic variance of the difference, and $v_{ik}^2(m)/m$ is the variance of the difference between batch means of size m from systems i and k.

If Assumption C holds, then $v_{ik}^2(m)/v_{ik}^2 = 1$ and the APICS is α. Since the two quantities typically are not equal we refer to $v_{ik}^2(m)/v_{ik}^2$ as the *bias*.

To facilitate the analysis, we use an AR(1) model as a surrogate for the simulation output process. Specifically,

$$X_{ij} = \mu_i + \phi(X_{i,j-1} - \mu_i) + Z_{i,j} \tag{4}$$

where $Z_{i,j} \overset{\text{iid}}{\sim} \text{N}(0, 1-\phi^2)$ and $-1 < \phi < 1$. For simplicity, suppose that each system has the same parameter ϕ. Under this assumption, the difference of two systems' output is also AR(1) with parameter, ϕ. We can show that the bias becomes

$$\frac{v_{ik}^2(m)}{v_{ik}^2} = 1 - \frac{2\phi}{(1-\phi^2)m} + \frac{2\phi^{m+1}}{(1-\phi^2)m} \tag{5}$$

so that the APICS is

$$\text{APICS} = 1 - \left\{1 - \frac{1}{2}\left[1 + 2\eta\left(1 - \frac{2\phi}{(1-\phi^2)m} + \frac{2\phi^{m+1}}{(1-\phi^2)m}\right)\right]^{-d/2}\right\}^{k-1}.$$

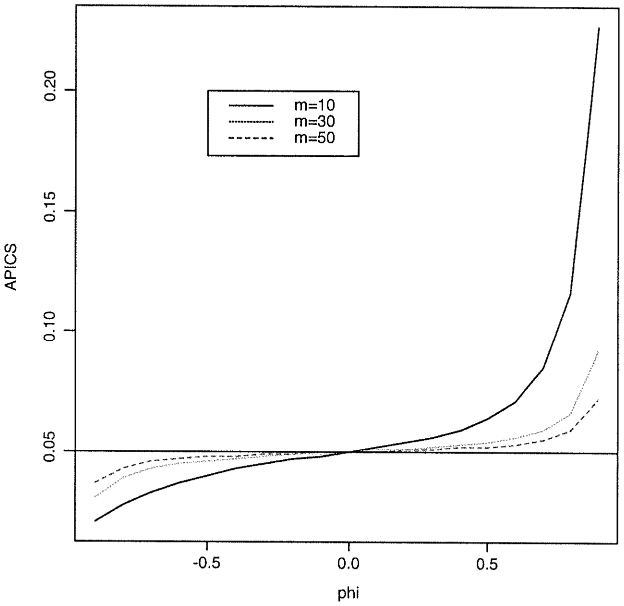

Figure 1 **APICS as a Function of ϕ for Different Batch Size (m) When $b = 5$, $k = 10$**

We computed the APICS for various values of m, b, k and ϕ, with η chosen as described in procedure $\mathcal{KN}+$. Although we considered the three variance estimators, BM, OBM and A, the analysis is the same for each of them except for the value of d. Therefore, we can judge the behavior of the APICS for the OBM or area estimators by looking at the APICS for BM only.

We varied m = 5, 10, 20 and b = 10, 30, 50, 80, while letting k and ϕ range over k = 2, 5, 10, 25, 100 and $\phi = -0.9, -0.8, \ldots, 0.9$. The nominal probability of correct selection was set at $1 - \alpha = 0.95$. The results are as follows:

- *Larger m implies closer-to-nominal APICS.* Not surprisingly, large batch size resulted in an APICS close to the nominal probability of incorrect selection. Figure 1 shows the APICS for different values of m when $b = 5$ and $k = 10$. In the case of $m = 10$, the APICS is much greater than 0.05 relative to when $m = 30$ or $m = 50$ for every positive value of ϕ. The same pattern appeared for different k and b. Larger batch size causes (5) to be closer to 1, making the APICS closer to the nominal level.
- *Larger k implies farther-from-nominal APICS.* Figure 2 shows the APICS for different k when $b = 5$ and $m = 10$. Notice that the APICS is farther from the nominal when $k = 100$ than it is when $k = 2$ or $k =$

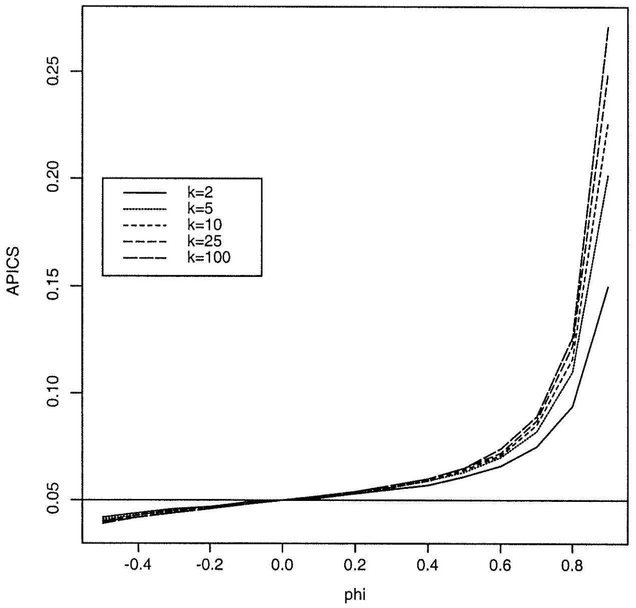

Figure 2 APICS as a Function of ϕ for a Different Number of Systems (k) When $b = 5$, $m = 10$

10 for all values of ϕ. This is because the effect of bias is amplified by the factor $k-1$ in the equation for APICS.

- *Negative ϕ always produces APICS smaller than* 0.05. On the other hand, positive ϕ makes the APICS bigger, thus requiring a larger batch size to achieve the nominal probability of incorrect selection.

These results suggest that one may want to pick m_0 as large as possible given n_0. When Goldsman and Marshall (1999) examined new two-stage procedures using standardized time series variance estimators, they showed that larger m_0 made the procedures less efficient, but more likely to deliver the desired PCS. This implies that large m_0 may be inefficient for $\mathcal{KN}+$, also. We examine this in our empirical analysis.

6. Empirical Analysis

In this section we report on a portion of an extensive empirical evaluation of $\mathcal{R}+$, $\mathcal{KN}+$ and $\mathcal{KN}++$. 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_0 (or batching sequence $\{(m_r, b_r)\}$); the configuration of the true means; and the dependence structure within the process. In all cases system 1 was the true best (had the largest or smallest mean, depending on the particular problem). 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 of equation (4); the moving average order-1, or MA(1), process

$$X_{ij} = \mu_i + \theta Z_{j-1} + Z_{i,j}$$

where $Z_{i,j} \overset{\text{iid}}{\sim} \mathrm{N}(0, 1/(1+\theta^2))$; and the waiting time process from an M/M/1 queue

$$X_{ij} = \max\{0, X_{i,j-1} + S_{i,j-1} - G_{ij}\}$$

where $S_{i,j-1} \overset{\text{iid}}{\sim} \exp(\tau_i)$ represents the service time, and $G_{ij} \overset{\text{iid}}{\sim} \exp(\lambda)$ represents the interarrival-time gap. The AR(1) and MA(1) processes are both marginally normal—which is favorable to our procedures—but the mean, variance and dependence can be controlled independently. The M/M/1 provides an example with non-normal marginals, but the mean, variance and dependence are all functions of the service rate τ and the arrival rate λ.

6.1. Configurations and Experiment Design

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

When we have independent data, $n_0 = 24$ 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 $v^2(n_0) = n_0 \mathrm{var}[\overline{X}(n_0)]$ and its limit, v^2, is approximately equal to 1; more specifically, $|1 - v^2(n_0)/v^2| \approx 0.01$. This guarantees that there is enough, but not too much, data so that it is possible to get a reasonably good estimator of v^2. After n_0 was determined (and it can be determined analytically for the AR(1) and MA(1) processes and empirically for M/M/1 processes), all divisors of n_0 were employed as batch sizes

m_0, implying n_0/m_0 batches for BM and A, and $n_0 - m_0 + 1$ for OBM.

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.

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

We now explain how we set the level of dependence and configurations of the means for each process.

6.1.1. AR(1) and MA(1) Processes. For the AR(1) and MA(1) processes, the strength of the correlation among the outputs depends on ϕ and θ, respectively. We varied ϕ and θ 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.

Two configurations of the true means were used: The first was 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. For AR(1) and MA(1) processes, the variances of all systems are the same.

6.1.2. M/M/1 Queue. The performance measure is w_i, the expected waiting time in the queue of system i. Thus, *smaller w_i is better* and system 1 is the best. In all cases the service rate of system 1 is set to $\tau_1 = 1$. To achieve different levels of dependence, the arrival rate varies over $\lambda = 0.3, 0.6, 0.9$ so the traffic intensity of system 1, $\rho = \lambda/\tau_1$, varies over 0.3, 0.6, 0.9.

For configurations of w_i, we consider SC and MDM configurations. In SC, $w_1 = \rho^2/\lambda(1-\rho)$, while $w_2 = \cdots = w_k = w_1 + \delta$. In MDM, $w_i = w_1 + (i-1)\delta$ so the means of all systems are δ apart. Table 1 shows one example of the MDM configuration. After we determine the desired w_i for $i = 2, 3, \ldots, k$, we derive the service rate τ_i that delivers it. We generate X_{ij} based on the algorithm of Schmeiser and Song (1989).

Table 1 **Example of the MDM Configuration for the M/M/1 Queue When $k = 3$**

	$\lambda = 0.3$	$\lambda = 0.6$	$\lambda = 0.9$
$w_1(\tau_1 = 1)$	0.429	1.5	9
w_2	$0.429+\delta$	$1.5+\delta$	$9+\delta$
w_3	$0.429+2\delta$	$1.5+2\delta$	$9+2\delta$

The MDM configuration is very interesting since a larger expected waiting time in the queue is associated with a larger variance. Thus, inferior systems have larger variances than the best system, and it is more difficult to find the best than it is when variances are the same across systems.

6.2. Batching Sequences

$\mathcal{KN}++$ requires a batching sequence (B) that guarantees the convergence in probability of the variance estimator; both strong consistency and mean-square-error (MSE) consistency imply convergence in probability. Damerdji (1994) showed that if $\{(m_r, b_r)\}$ is a batching sequence satisfying certain conditions, then variance estimators become strongly consistent as m_r and b_r go to infinity. Mean-square consistency is also studied in Damerdji (1995) and Damerdji and Goldsman (1995) for various variance estimators. Song and Schmeiser (1995) and Chien et al. (1997) derive the optimal mean-squared-error batch size.

In this section, we review several batching sequences that have been proposed in the literature, and a modification of them that we propose. In Section 6.3 we will compare, empirically, the performance of the batching sequences in terms of PCS and total number of observations.

In our procedure, we start with a certain first-stage sample size, n_0, divided into b_0 batches of size m_0; that is, $n_0 = m_0 b_0$. Three different batching sequences that have been proposed are the following:

- B_1: $m_r = \sqrt[3]{r}$. Song and Schmeiser (1995) showed that this rule leads to the smallest asymptotic MSE.
- B_2: $m_r = \sqrt{r}$. This is often called the square-root rule (Fishman and Yarberry 1997). The square-root rule is known to achieve the fastest convergence rate to the asymptotic variance, but can produce low confidence-interval coverage.
- B_3: $m_r = \sqrt[3]{r^2}$. Damerdji and Goldsman (1995) showed that the batch size m should be increased

faster than $\sqrt{r}$ to achieve strong consistency of the variance estimator.
Thus, we picked $\sqrt[3]{r^2}$.

If, in the first stage, we start with a large value of the batch size m_0 (small number of batches b_0), then it may be a very long time (measured in number of observations) until an update occurs. For example, if $n_0 = 100$ and $m_0 = 50$, then the first update will occur at $r = 125{,}000$ for B_1 and B_3, and at $r = 2{,}500$ for B_2. Thus, the update may never happen and we will not realize the benefits of updating. On the other hand, we do not want to risk having m_0 too small since a severely biased variance estimator may lead to early and incorrect termination of the procedure (when output data are positively correlated the variance estimators tend to be biased low).

To induce more frequent updates, we introduce a sequence that doubles the number of batches (for BM and A), while fixing the batch size at m_0, until B_1, B_2 or B_3 can be applied. For OBM, the number of batches increases at those points when the total sample size doubles. To express this in an algorithm format, let

$$u_\ell(r) = \begin{cases} \sqrt[3]{r}, & \ell = 1 \\ \sqrt{r}, & \ell = 2 \\ \sqrt[3]{r^2}, & \ell = 3. \end{cases}$$

Then our modification of B_ℓ is as follows:

Modified Batching Sequence

```
Setup: Pick $n_0$, $m_0$ and $b_0$ such that $n_0 = b_0 m_0$ and
    set $r = n_0 + 1$, $m_r = m_0$, $b_r = b_0$ and $f = 2$.
Update: For each new value of $r$
    If ($u_\ell(r) < m_0$ & $r = f n_0$) then
        Set $m_r \leftarrow m_{r-1}$
        Set $b_r \leftarrow 2b_{r-1}$ for BM and A
        Set $b_r \leftarrow r - m_r + 1$ for OBM
        Set $f \leftarrow 2f$
    ElseIf ($u_\ell(r) \geq m_0$ & $u_\ell(r)$ is integer) then
        Set $m_r \leftarrow u_\ell(r)$
        Set $b_r \leftarrow \lfloor r/m_r \rfloor$ for BM and A
        Set $b_r \leftarrow r - m_r + 1$ for OBM
    Else
        Set $m_r \leftarrow m_{r-1}$
        Set $b_r \leftarrow b_{r-1}$
    Endif
```

Table 2 Modified Batching Sequence for BM and A When $n_0 = 100$ and $m_0 = 50$ (Table Entries Are Given Only Where the Batch Size Changes)

r	B_1	B_2	B_3
100	(50, 2)	(50, 2)	(50, 2)
200	(50, 4)	(50, 4)	(50, 4)
400	(50, 8)	(50, 8)	(50, 8)
800	(50, 16)	(50, 16)	(50, 16)
1,600	(50, 32)	(50, 32)	(50, 32)
2,500		(50, 50)	
2,601	⋮	(51, 51)	⋮
2,704		(52, 52)	
3,200	(50, 64)	(56, 56)	(50, 64)
⋮	⋮	⋮	⋮

Continuing with the same example as above, the modified batching sequences for BM and A appear in Table 2.

6.3. Summary of Results

The experiments show that $\mathcal{KN}+$ and $\mathcal{KN}++$ are typically more efficient than $\mathcal{R}+$. The gain in efficiency is due to the ability of $\mathcal{KN}+$ and $\mathcal{KN}++$ to eliminate inferior systems. The performance of $\mathcal{KN}++$ is particularly excellent, with savings in the total number of observations of up to 99% relative to $\mathcal{R}+$.

To illustrate the key conclusions, we report selected results, emphasizing cases when strong dependence exists (ϕ, θ and ρ equal to 0.9) and the means are arrayed in the MDM configuration. Tables 3–4 show

Table 3 Sample Average of Total Basic (Unbatched) Observations When AR(1) Processes Are Tested with the MDM Configuration, $k = 10$, $\phi = 0.9$ and $n_0 = 1000$ (All Values Are in Units of 10^4)

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000			47.63			1667.83			
500	47.65	46.52	42.38	1686.39	39.45	38.40	4.36	4.40	4.07
250	34.51	23.95	28.53	12.86	8.03	8.21	3.38	3.45	3.32
200	28.43	21.18	23.94	7.85	6.07	6.03	3.18	3.22	3.07
125	19.04	15.71	16.38	4.44	4.07	3.85	3.91	3.75	3.49
100	16.54	14.42	13.98	3.75	3.56	3.13	3.37	3.30	2.95
50	12.12	11.44	7.82	2.63	2.61	1.83	2.29	2.36	1.69
40	11.06	10.58	6.26	2.39	2.38	1.54	2.25	2.32	1.52

Table 4 Estimated PCS When AR(1) Processes Are Tested with the MDM Configuration, $k = 10$, $\phi = 0.9$ and $n_0 = 1000$ (No Values Are Statistically Significantly Smaller than 0.95)

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
1000			0.994			0.993			
500	0.995	1.000	0.994	0.994	1.000	0.997	0.985	0.997	0.983
250	0.992	0.995	0.994	0.990	0.997	0.995	0.986	0.992	0.987
200	0.992	0.996	0.993	0.996	0.998	0.993	0.984	0.991	0.983
125	0.990	0.992	0.991	0.991	0.995	0.989	0.997	0.999	0.989
100	0.993	0.991	0.988	0.995	0.994	0.989	0.992	0.996	0.983
50	0.988	0.986	0.967	0.988	0.989	0.962	0.990	0.991	0.969
40	0.988	0.986	0.950	0.982	0.984	0.946	0.986	0.988	0.946

the sample average of the total number of basic (unbatched) observations (OBS) and estimated PCS for the AR(1) processes when BM, OBM and A are employed and $k = 10$. By "total number of basic (unbatched) observations" we mean all observations, including the initial n_0, generated by the simulation until the selection procedure terminates. Tables 5–6 show OBS and estimated PCS for the MA(1) processes when $k = 10$. Tables 7–8 report OBS and estimated PCS for the M/M/1 processes when there are $k = 5$ systems. All of the results reported for $\mathcal{KN}++$ employ B_2 as the batching sequence. Notice that since we fix the first-stage sample size, n_0, large m_0 implies small b_0.

- **Effect of m_0 and b_0**: Tables 3, 5 and 7 show that each procedure consumes a very large number of observations when m_0 is large (b_0 small). Both $\mathcal{R}+$ and $\mathcal{KN}+$, in particular, are inefficient when they start with the largest possible m_0 for a given n_0. For the efficiency of the procedures, small m_0 is desirable.

Table 5 Sample Average of Total Basic (Unbatched) Observations When MA(1) Processes Are Tested with the MDM Configuration, $k = 10$, $\theta = 0.9$ and $n_0 = 60$ (All Values Are in Units of 10^4)

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
60			40.35			99.69			
30	40.27	8.71	8.99	99.07	2.44	2.41	0.78	0.73	0.59
20	8.76	3.03	3.11	2.40	0.81	0.83	0.34	0.33	0.30
15	3.03	1.53	1.98	0.80	0.49	0.50	0.23	0.23	0.21
12	1.93	1.31	1.55	0.49	0.38	0.38	0.19	0.20	0.18
10	1.51	1.10	1.32	0.37	0.31	0.31	0.18	0.19	0.17

Table 6 Estimated PCS When MA(1) Processes Are Tested with the MDM Configuration, $k = 10$, $\theta = 0.9$ and $n_0 = 60$ (No Values Are Statistically Significantly Smaller than 0.95)

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
60			0.995			0.996			
30	0.995	1.000	0.994	0.995	1.000	0.997	0.995	0.999	0.994
20	0.994	0.999	0.993	0.996	0.998	0.995	0.990	0.997	0.995
15	0.993	0.995	0.994	0.996	0.998	0.990	0.983	0.994	0.986
12	0.993	0.995	0.994	0.996	0.997	0.994	0.987	0.996	0.988
10	0.994	0.993	0.994	0.993	0.997	0.995	0.987	0.996	0.988

On the other hand, the estimated PCS decreases as m_0 decreases as seen in Tables 4, 6 and 8. The numbers reported in these three tables for estimated PCS are greater than the nominal PCS of 0.95 when we have large m_0. This implies that large m_0 helps to achieve the nominal PCS but at the cost of a huge number of observations.

- **Performance of $\mathcal{R}+$, $\mathcal{KN}+$ and $\mathcal{KN}++$**: Tables 3, 5 and 7 show that $\mathcal{R}+$ outperforms $\mathcal{KN}+$ only in the extreme case when each variance estimator has 1 degree of freedom. For instance, in Table 3, A and BM have 1 degree of freedom when $m_0 = 1000$ and $m_0 = 500$, respectively, and $\mathcal{R}+$ is more efficient than $\mathcal{KN}+$. In this extreme case, there is no benefit from using $\mathcal{KN}+$ because it loses its ability to quickly eliminate inferior systems. However, the performance of $\mathcal{KN}++$ is not affected as much by the choice of m_0 or b_0, and it is more efficient than the other two procedures. The reason is that $\mathcal{KN}++$ can correct a poor initial variance estimate by updating it as more data are obtained. The disadvantage of $\mathcal{KN}++$ is that it requires saving all of the data in order to compute the variance updates. In addition, when the SC configuration is tested, $\mathcal{KN}++$ shows some coverage problems when m_0 is small. However, as long as $\mathcal{KN}++$ starts with the largest possible value of m_0 for given n_0, the estimated PCS is reasonably good while the procedure remains efficient.
- **Variance Estimators**: For a given initial batch size m_0, procedures based on OBM and A require fewer observations than those based on BM. One reason is that OBM and A have greater degrees of freedom than BM for a given m_0. BM and OBM have about the same bias as estimators of the asymptotic variance, but for large batch size and num-

Table 7 **Sample Average of Total Basic (Unbatched) Observations When M/M/1 Processes Are Tested with the MDM Configuration, $k = 5$, $\rho = \lambda/\tau_1 = 0.9$ and $n_0 = 24000$ (All Values Are in Units of 10^5)**

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
24000			389.23			1,387.83			
12000	391.24	158.52	150.61	1,329.52	65.49	67.85	7.89	7.99	7.44
8000	154.15	82.62	81.59	66.38	27.27	27.74	6.91	6.98	6.61
6000	82.40	50.97	57.49	26.68	18.03	17.60	6.48	6.51	6.19
4800	60.04	44.70	46.48	18.04	14.10	13.24	6.04	6.05	5.68
4000	50.18	38.81	40.12	14.22	12.00	10.82	5.85	5.85	5.36
3000	39.58	33.27	30.90	10.44	9.73	8.36	5.44	5.43	4.80
2400	35.32	31.31	25.79	8.94	8.48	6.78	5.15	5.12	4.25
2000	32.39	28.69	21.59	8.00	7.70	5.76	4.92	4.93	3.99
1600	28.86	26.20	17.07	7.07	6.86	4.60	4.70	4.72	3.51
1000	22.76	21.48	10.51	5.48	5.40	2.82	4.11	4.11	2.49

ber of batches OBM's variance is about 1/3 smaller than BM (Song and Schmeiser 1995). And while A is first-order unbiased, the variances of BM and A are about the same. Procedures based on OBM typically achieve the highest estimated PCS among all variance estimators.

- **SC vs. MDM**: Table 9 shows the effect of having different configurations of the true means. When AR(1) and MA(1) processes are tested, $\mathcal{R}+$ consumes about the same number of observations regardless of whether the configuration is SC or MDM, because the variances in the AR(1) and MA(1) examples are not affected by a shift in the means and N_i is determined by the variance estimator of system i. In the M/M/1 queue, however, $\mathcal{R}+$ requires more observations in the MDM configuration than the SC configuration because variances do depend on the means and, for this example, the inferior systems have larger variances in the MDM configuration. $\mathcal{KN}+$ and $\mathcal{KN}++$ always consume fewer observations when the MDM configuration is employed rather than when SC is in force; in the MDM configuration inferior systems are easier to eliminate. Table 9 shows that when AR(1) and MA(1) processes are tested the savings is as large as 50%. When M/M/1 processes are tested the savings is less dramatic, but still substantial.

Table 8 **Estimated PCS When M/M/1 Processes Are Tested with the MDM Configuration, $k = 5$, $\rho = \lambda/\tau_1 = 0.9$ and $n_0 = 24000$ (Values That Are Statistically Smaller than 0.95 Are Enclosed in a □)**

	$\mathcal{R}+$			$\mathcal{KN}+$			$\mathcal{KN}++$		
m_0	BM	OBM	A	BM	OBM	A	BM	OBM	A
24000			0.986			0.987			
12000	0.984	0.994	0.971	0.979	0.997	0.979	0.975	0.982	0.959
8000	0.976	0.982	0.957	0.992	0.990	0.979	0.970	0.987	0.967
6000	0.968	0.967	0.962	0.985	0.984	0.974	0.961	0.979	0.967
4800	0.970	0.959	0.952	0.978	0.981	0.972	0.958	0.965	0.955
4000	0.957	0.956	0.951	0.974	0.975	0.956	0.966	0.971	0.959
3000	0.956	0.950	0.949	0.965	0.969	0.962	0.967	0.972	0.957
2400	0.953	0.948	0.940	0.961	0.970	0.959	0.971	0.971	0.960
2000	0.956	0.945	$\boxed{0.934}$	0.965	0.965	0.941	0.963	0.970	0.943
1600	0.949	0.941	$\boxed{0.934}$	0.959	0.958	0.942	0.948	0.946	$\boxed{0.915}$
1000	$\boxed{0.936}$	0.938	$\boxed{0.911}$	0.946	0.950	$\boxed{0.896}$	0.959	0.958	$\boxed{0.909}$

Table 9 **The Effect of SC vs. MDM Configurations on Sample Average of Total Basic (Unbatched) Observations When Estimator A is Used (All Values Are in Units of 10^4)**

			$\mathcal{R}+$	$\mathcal{KN}+$	$\mathcal{KN}++$
AR(1)	$k = 10$, $\phi = 0.9$	SC	23.94	13.63	5.98
	$m_0 = 200$	MDM	23.94	6.03	3.07
M/M/1	$k = 5$, $\rho = 0.9$	SC	291.91	178.78	66.92
	$m_0 = 4800$	MDM	464.78	132.36	56.76

- **Batching Sequence for $\mathcal{KN}++$:** Table 10 reports OBS and estimated PCS when M/M/1 processes are tested with the OBM estimator, $\rho = 0.6$, $k = 10$ and $n_0 = 1200$. When $\mathcal{KN}++$ starts with a very large initial batch size m_0, the procedure terminates during the phase in which b_i is doubled, but before ever reaching the actual batching sequence, $u_\ell(r)$. When the procedure starts with a smaller batch size m_0 ($m_0 \leq 200$ in Table 10), $u_\ell(r)$ is invoked. Table 10 shows that up to some point $\mathcal{KN}++$ using B_2 consumes a slightly larger number of observations, but has higher estimated PCS, than B_1 or B_3, but thereafter it consumes fewer observations and has lower estimated PCS. There seems to be no difference in performance between B_1 and B_3.

For $\mathcal{R}+$ and $\mathcal{KN}+$, a strategy for selecting m_0 and b_0 is not clear. Assuming that one wants to find the best among a very large number of alternatives, say larger than 20, we suggest that any m_0 which yields $b_0 \geq 5$ can be chosen. For $\mathcal{KN}++$, we strongly recommend that one choose the largest possible m_0 given n_0 to help the procedure to achieve the nominal PCS.

Table 10 **The Effect of Different Batching Sequences on $\mathcal{KN}++$ When M/M/1 Processes Are Tested with the SC Configuration, $n_0 = 1200$, $k = 10$, $\rho = 0.6$ and the OBM Estimator (Values of Total Basic [Unbatched] Observations Are in Units of 10^4)**

	Total OBS			Estimated PCS		
m_0	B_1	B_2	B_3	B_1	B_2	B_3
600	10.75	10.75	10.75	0.952	0.952	0.952
400	9.54	9.54	9.54	0.939	0.939	0.939
300	8.64	8.64	8.64	0.926	0.926	0.926
240	8.55	8.55	8.55	0.927	0.927	0.927
200	7.97	12.29	7.97	0.899	0.921	0.899
150	7.60	11.47	7.60	0.896	0.916	0.896
120	7.23	8.98	7.23	0.868	0.875	0.868
100	6.96	7.52	6.96	0.890	0.868	0.890
80	7.08	6.63	7.08	0.887	0.869	0.887
75	7.03	6.54	7.03	0.845	0.842	0.845

7. The Future

The empirical evidence presented here, as well as other analyses 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 $\mathcal{R}+$ 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. $\mathcal{KN}+$ and $\mathcal{KN}++$ are highly efficient procedures, but they assume the ability to obtain incremental output data from each system in a coordinated manner.

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 $\mathcal{R}+$ and $\mathcal{KN}+$, 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 $\mathcal{KN}++$ 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 (2001a) for $\mathcal{KN}$. Because CRN induces dependence across systems, and we already have dependence within replications, it becomes difficult to provide procedures that account for both.

Acknowledgments

This research was partially supported by National Science Foundation Grant DMI-9622065 and DMI-9622269 and by Symix Corporation/Pritsker Division and Rockwell Software. Portions of this paper were previously published in Goldsman, Kim, Marshall, and

Nelson (2000). The authors thank the Associate Editor and referee for their help in revising the paper.

References

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

Chien, C., D. Goldsman, B. Melamed. 1997. Large-sample results for batch means. *Management Science* **43** 1288–1295.

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

Damerdji, H. 1995. Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Operations Research* **43** 282–291.

Damerdji, H., D. Goldsman. 1995. Consistency of several variants of the standardized time series area variance estimator. *Naval Research Logistics* **42** 1161–1176.

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

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

Dudewicz, E. J., 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.

Fishman, G. S., L. S. Yarberry. 1997. An implementation of the batch means method. *INFORMS Journals on Computing* **9** 296–310.

Glynn, P. W., W. Whitt. 1991. Estimating the asymptotic variance with batch means. *Operations Research Letters* **10** 431–435.

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

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

Goldsman, D., S.-H. Kim, W. S. Marshall, B. L. Nelson. 2000. Ranking and selection for steady-state simulation. J. Joines, R. R. Barton, P. Fishwick, K. Kang, eds. *Proceedings of the 2000 Winter Simulation Conference.* Institute of Electrical and Electronics Engineers, Piscataway, NJ. 544–553.

Goldsman, D., W. S. Marshall. 1999. Selection procedures with standardized time series variance estimators. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, G. W. Evans, eds. *Proceedings of the 1999 Winter Simulation Conference.* Institute of Electrical and Electronics Engineers, Piscataway, NJ. 382–388.

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

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

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

Kim, S.-H., B. L. Nelson. 2001a. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS,* in press.

Kim, S.-H., B. L. Nelson. 2001b. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. Working Paper, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL.

Law, A. M., W. D. Kelton. 2000. *Simulation Modeling & Analysis,* 3rd edition. McGraw-Hill, New York.

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

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

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.

Schmeiser, B. W., W. T. Song. 1989. Inverse-transformation algorithms for some common stochastic processes. E. A. MacNair, K. J. Musselman, P. Heidelberger, eds. *Proceedings of the 1989 Winter Simulation Conference.* Institute of Electrical and Electronics Engineers, Piscataway, NJ. 490–496.

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

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

Steiger, N. M., J. R. Wilson. 2001. An improved batch means procedure for simulation output analysis. *INFORMS Journal on Computing,* in press.

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

Accepted by W. David Kelton; received March 2001; revised September 2001; accepted September 2001.