A Fully Sequential Procedure for Indifference-Zone Selection in Simulation

SEONG-HEE KIM
Georgia Institute of Technology
and
BARRY L. NELSON
Northwestern University

We present procedures for selecting the best or near-best of a finite number of simulated systems when best is defined by maximum or minimum expected performance. The procedures are appropriate when it is possible to repeatedly obtain small, incremental samples from each simulated system. The goal of such a sequential procedure is to eliminate, at an early stage of experimentation, those simulated systems that are apparently inferior, and thereby reduce the overall computational effort required to find the best. The procedures we present accommodate unequal variances across systems and the use of common random numbers. However, they are based on the assumption of normally distributed data, so we analyze the impact of batching (to achieve approximate normality or independence) on the performance of the procedures. Comparisons with some existing indifference-zone procedures are also provided.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: multivariate statistics; I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis

General Terms: Experimentation, Theory

Additional Key Words and Phrases: Multiple comparisons, output analysis, ranking and selection, variance reduction

1. INTRODUCTION

In a series of papers [Boesel et al. 2001; Goldsman and Nelson 1998a; 1998b; Nelson and Banerjee 1999; 2001; Nelson and Goldsman 2001; Nelson et al. 2001; Miller et al. 1996; 1998a; 1998b], we have addressed the problem of selecting the best simulated system when the number of systems is finite and no functional relationship among the systems is assumed. We have focused primarily

This work was supported by the National Science Foundation under Grant Number DMI-9622065, Rockwell Software and Symix/Pritsker.

Authors' addresses: S.-H. Kim, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332; e-mail: skim@isye.gatech.edu; B. L. Nelson, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3119; e-mail: nelsonb@northwestern.edu.

on situations in which "best" is defined by maximum or minimum expected performance, which is also the definition we adopt in the present article,

Our work grows out of the substantial literature on ranking, selection, and multiple comparison procedures in statistics (see, for instance, Bechhofer et al. [1995], Hochberg and Tamhane [1987], and Hsu [1996]), particularly the "indifference zone" approach in which the experimenter specifies a practically significant difference worth detecting. Our approach has been to adapt, extend and invent procedures to account for situations and opportunities that are common in simulation experiments, but perhaps less so in physical experiments. These include:

—Unknown and unequal variances across different simulated systems.

—Dependence across systems' outputs due to the use of common random numbers.

—Dependence within a system's output when only a single replication is obtained from each system in a "steady-state simulation."

—A very large number of alternatives that differ widely in performance.

—Alternatives that are available sequentially or in groups, rather than all at once, as might occur in an exploratory study or within an optimization/search procedure.

Prior to the present article we have proposed procedures that kept the number of stages small, say 1, 2 or 3, where a "stage" occurs whenever we initiate a simulation of a system to obtain data. It makes sense to keep the number of stages small when they are implemented manually by the experimenter, or when it is difficult to stop and restart simulations. However, as simulation software makes better use of modern computing environments, the programming difficulties in switching among alternatives to obtain increments of data are diminishing (although there may still be substantial computing overhead incurred in making the switch). The procedures presented in this paper can, if desired, take only a single basic observation from each alternative still in play at each stage. For that reason, they are said to be "fully sequential with elimination." The motivation for adopting fully sequential procedures is to reduce the overall simulation effort required to find the best system by eliminating apparently inferior alternatives early in the experimentation. Of course, there are 2- or 3-stage procedures with elimination. However, fully sequential procedures have many opportunities to discard inferior systems, systems that might not be detected by 2- or 3-stage procedures until the final stage. Thus, fully sequential procedures are expected to be more efficient than other competitors in the sense that fewer observations and less computer time are needed to find the best.

For those situations in which there is substantial computing overhead when switching among alternative systems, we also evaluate the benefits of taking batches of data—rather than a single observation—from each system at each stage. These results have implications for the steady-state simulation problem when the method of batch means is employed, or when the simulation data are not approximately normally distributed.

Our work can be viewed as extending, in several directions, the results of Paulson [1964] and Hartmann [1991], specifically in dealing with unequal variances across systems and dependence across systems due to the use of common random numbers (CRN) (see also Hartmann [1988], Bechhofer et al. [1990], and Jennison et al. [1982]). Chick and Inoue [2001a, 2001b] also present procedures that seek to efficiently allocate observations so as to find the best system. Their approach takes a Bayesian perspective that maximizes the experimenter's posterior probability of a correct selection given a computing budget. (See also Chen [1996]).

The article is organized as follows: In Section 2, we provide an algorithm for our fully sequential procedure and prove its validity, by which we mean proving that a prespecified probability of correct selection is attained. Section 3 provides guidance on how to choose various design parameters of the procedure, including critical constants, batch size and whether or not to use CRN. Some empirical results are provided in Section 4, followed by conclusions in Section 5.

2. THE PROCEDURE

In this section, we describe a fully sequential procedure that guarantees, with confidence level greater than or equal to $1 - \alpha$, that the system ultimately selected has the largest true mean when the true mean of the best 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 procedure guarantees to find one of these "good" systems with the same probability. The parameter δ, which is termed the *indifference zone*, is set by the experimenter to the smallest actual difference that it is important to detect. Differences of less than δ are considered practically insignificant.

The procedure is sequential, has the potential to eliminate alternatives from further consideration at each stage, and terminates with only one system remaining in contention. However, the experimenter may also choose to terminate the procedure when there are $m \geq 1$ systems still in contention, in which case the procedure guarantees that the subset of size m contains the best system (or one of the good systems) with confidence greater than or equal to $1 - \alpha$.

Throughout the article, we use the notation X_{ij} to indicate the jth independent observation from system i. We assume that the $X_{ij} \sim \mathrm{N}(\mu_i, \sigma_i^2)$, with both μ_i and σ_i^2 unknown. Notice that X_{ij} may be the mean of a batch of observations that are also individually normal, provided the batch size remains fixed throughout the procedure (we analyze the effect of batch size later). We 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. The procedure is valid with or without the use of common random numbers.

Fully Sequential, Indifference-Zone Procedure

Setup: Select confidence level $1-\alpha$, indifference zone δ and first-stage sample size $n_0 \geq 2$. Calculate η and c as described below.

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

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

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

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 $n_0 > \max_i N_i$, 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

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

Notice that $W_{i\ell}(r)$, which determines how far the sample mean from system i can drop below the sample means of the other systems without being eliminated, decreases monotonically as the number of replications r increases.

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 the stopping rule can also be $|I| = m > 1$ if it is desired to find a subset containing the best, rather than the single best.)

Constants: The constant c may be any nonnegative integer, with standard choices being $c = 1, 2$; these values are standard in the sense that they were used by Hartmann [1991], and that η is easy to compute when $c = 1$ or 2. We evaluate different choices later in the paper and argue that $c = 1$ is typically the best choice. The constant η is the solution to the equation

$$\begin{aligned} g(\eta) &\equiv \sum_{\ell=1}^{c} (-1)^{\ell+1} \left(1 - \frac{1}{2} \mathcal{I}(\ell = c) \right) \left(1 + \frac{2\eta(2c - \ell)\ell}{c} \right)^{-(n_0-1)/2} \\ &= \frac{\alpha}{k - 1} \end{aligned} \qquad (1)$$

where $\mathcal{I}$ is the indicator function. In the special case that $c = 1$, we have the closed-form solution

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

To prove the validity of the procedure, we need the following lemmas from Fabian [1974] and Tamhane [1977]:

LEMMA 1 (FABIAN [1974]). *Let $X_1, X_2, \ldots$ be independent and identically distributed* $\mathrm{N}(\Delta, 1)$ *random variables with* $\Delta > 0$. *Let*

$$\begin{aligned} S(n) &= \sum_{j=1}^{n} X_j \\ L(n) &= -a + \gamma n \\ U(n) &= a - \gamma n \end{aligned}$$

for some $a > 0$ and $\gamma \geq 0$. Let $R(n)$ denote the interval $[L(n), U(n)]$, and let $T = \min\{n : S(n) \notin R(n)\}$ be the first time the partial sum $S(n)$ does not fall in the triangular region defined by $R(n)$. We assume that $R(n) = \emptyset$ when $L(n) > U(n)$. Finally, let $\mathcal{E}$ be the event $\{S(T) \leq L(T)$ and $R(T) \neq \emptyset$, or $S(T) \leq 0$ and $R(T) = \emptyset\}$. If $\gamma = \Delta/(2c)$ for some positive integer c, then

$$\Pr\{\mathcal{E}\} \leq \sum_{\ell=1}^{c} (-1)^{\ell+1} \left(1 - \frac{1}{2}\mathcal{I}(\ell = c)\right) \exp\{-2a\gamma(2c - \ell)\ell\}.$$

Remark 1. In our proof that the fully sequential procedure provides the stated correct selection guarantee, the event $\mathcal{E}$ will correspond to an incorrect selection (incorrectly eliminating the best system from consideration).

LEMMA 2 (TAMHANE [1977]). *Let $V_1, V_2, \ldots, V_k$ be independent random variables, and let $g_j(v_1, v_2, \ldots, v_k)$, $j = 1, 2, \ldots, p$, be nonnegative, real-valued functions, each one nondecreasing in each of its arguments. Then*

$$\mathrm{E}\left[\prod_{j=1}^{p} g_j(V_1, V_2, \ldots, V_k)\right] \geq \prod_{j=1}^{p} \mathrm{E}[g_j(V_1, V_2, \ldots, V_k)].$$

Without loss of generality, suppose that the true means of the systems are indexed so that $\mu_k \geq \mu_{k-1} \geq \cdots \geq \mu_1$ and let $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})'$ be a vector of observations across all k systems.

THEOREM 1. *Suppose that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent and identically distributed multivariate normal with unknown mean vector $\boldsymbol{\mu}$, that is arbitrary except for the condition that $\mu_k \geq \mu_{k-1} + \delta$, and unknown and arbitrary positive definite covariance matrix Σ. Then with probability $\geq 1 - \alpha$ the fully sequential indifference-zone procedure selects system k.*

PROOF. We begin by considering the case of only two systems, denoted k and i, with $\mu_k \geq \mu_i + \delta$. Select a value of η such that $g(\eta) = \beta$ for some $0 < \beta < 1/2$. Let

$$T = \min\left\{r : r \geq n_0 \text{ and } -W_{ik}(r) < \bar{X}_k(r) - \bar{X}_i(r) < W_{ik}(r) \text{ is violated}\right\}. \quad (3)$$

Notice that T is the stage at which the procedure terminates. Let ICS denote the event that an incorrect selection is made at time T. Then

$$
\begin{aligned}
\Pr\{\text{ICS}\} &= \Pr\{\bar{X}_k(T) - \bar{X}_i(T) < -W_{ik}(T)\} \\
&\le \Pr_{\text{SC}}\{\bar{X}_k(T) - \bar{X}_i(T) < -W_{ik}(T)\} \\
&= \Pr_{\text{SC}}\left\{\sum_{j=1}^{T}(X_{kj} - X_{ij}) \le \min\left\{0, \frac{-h^2 S_{ik}^2}{\delta 2c} + \frac{\delta T}{2c}\right\}\right\} \\
&= \Pr_{\text{SC}}\left\{\sum_{j=1}^{T}\left(\frac{X_{kj} - X_{ij}}{\sigma_{ik}}\right) \le \min\left\{0, \frac{-h^2 S_{ik}^2}{\delta 2c\sigma_{ik}} + \frac{\delta T}{2c\sigma_{ik}}\right\}\right\} \\
&= \mathrm{E}\left[\Pr_{\text{SC}}\left\{\sum_{j=1}^{T}\left(\frac{X_{kj} - X_{ij}}{\sigma_{ik}}\right) \le \min\left\{0, \frac{-h^2 S_{ik}^2}{\delta 2c\sigma_{ik}} + \frac{\delta T}{2c\sigma_{ik}}\right\} \middle| S_{ik}\right\}\right], \quad (4)
\end{aligned}
$$

where $\sigma_{ik}^2 = \mathrm{Var}[X_{kj} - X_{ij}]$ and "SC" denotes the slippage configuration $\mu_k = \mu_i + \delta$.

Notice that under the SC, $(X_{kj} - X_{ij})/\sigma_{ik}$ are independent and identically distributed $\mathrm{N}(\Delta, 1)$ with $\Delta = \delta/\sigma_{ik}$. In Lemma 1, let

$$
a = \frac{h^2 S_{ik}^2}{\delta 2c\sigma_{ik}} = \frac{\eta(n_0 - 1)S_{ik}^2}{\delta\sigma_{ik}}
$$

and $\gamma = \delta/(2c\sigma_{ik}) = \Delta/(2c)$. Therefore, the lemma implies that

$$
\begin{aligned}
&\mathrm{E}\left[\Pr_{\text{SC}}\left\{\sum_{j=1}^{T}\left(\frac{X_{kj} - X_{ij}}{\sigma_{ik}}\right) \le \min\left\{0, \frac{-h^2 S_{ik}^2}{\delta 2c\sigma_{ik}} + \frac{\delta T}{2c\sigma_{ik}}\right\} \middle| S_{ik}\right\}\right] \\
&\quad \le \mathrm{E}\left[\sum_{\ell=1}^{c}(-1)^{\ell+1}\left(1 - \frac{1}{2}\mathcal{I}(\ell = c)\right)\exp\{-2a\gamma(2c - \ell)\ell\}\right]. \quad (5)
\end{aligned}
$$

But observe that

$$
-2a\gamma(2c - \ell)\ell = -\frac{\eta(2c - \ell)\ell}{c} \times \frac{(n_0 - 1)S_{ik}^2}{\sigma_{ik}^2}
$$

and $(n_0 - 1)S_{ik}^2/\sigma_{ik}^2$ has a chi-squared distribution with $n_0 - 1$ degrees of freedom. To evaluate the expectation in (5), recall that $\mathrm{E}[\exp\{t\chi_\nu^2\}] = (1 - 2t)^{-\nu/2}$ for $t < 1/2$ and χ_ν^2 a chi-squared random variable with ν degrees of freedom (this expectation is the moment generating function of a chi-squared random variable). Thus, the expected value of (5) is

$$
\sum_{\ell=1}^{c}(-1)^{\ell+1}\left(1 - \frac{1}{2}\mathcal{I}(\ell = c)\right)\left(1 + \frac{2\eta(2c - \ell)\ell}{c}\right)^{-(n_0-1)/2} = \beta,
$$

where the equality follows from the way we choose η.

Thus, we have a bound on the probability of an incorrect selection when there are two systems. Now consider $k \ge 2$ systems, set $\beta = \alpha/(k - 1)$, and let ICS_i be the event that an incorrect selection is made when systems k and i are considered in isolation. Then

$$\Pr\{\text{ICS}\} \le \sum_{i=1}^{k-1} \Pr\{\text{ICS}_{\text{i}}\} \le \sum_{i=1}^{k-1} \frac{\alpha}{k-1} = \alpha$$

where the first inequality follows from the Bonferroni inequality. □

Remark 2. This procedure is valid with or without the use of common random numbers, since the effect of CRN is to change (ideally reduce) the value of σ_{ik}^2, which is not important in the proof. Notice that reducing σ_{ik}^2 will tend to reduce S_{ik}^2, which narrows (by decreasing a) and shortens (by decreasing N_i) the continuation region $R(n)$. Thus, CRN should allow alternatives to be eliminated earlier in the sampling process.

Corollary 1. *If $\mu_k < \mu_{k-1} + \delta$, then with probability $\ge 1 - \alpha$ the fully sequential indifference-zone procedure selects one of the systems whose mean is within δ of μ_k.*

Proof. If $\mu_k < \mu_1 + \delta$, then the result is trivially true, since any selected system constitutes a correct selection.

Suppose there exists $t > 1$ such that $\mu_k < \mu_t + \delta$, but $\mu_k \ge \mu_{t-1} + \delta$, and let CS denote the event that a correct selection is made when the procedure is applied to all k systems. Then

$$\begin{aligned}
\Pr\{\text{CS}\} &= \Pr\{\text{systems } 1, 2, \ldots, t-1 \text{ are eliminated}\} \\
&\ge \Pr\{\text{system } k \text{ eliminates } 1, 2, \ldots, t-1\} \\
&= 1 - \Pr\{\text{ICS; one of } 1, 2, \ldots, t-1 \text{ eliminates } k\} \\
&\ge 1 - \sum_{i=1}^{t-1} \Pr\{\text{ICS}_i\} \\
&\ge 1 - \frac{t-1}{k-1}\alpha \ge 1 - \alpha.
\end{aligned}$$

The first inequality follows because it is more difficult for system k alone to eliminate systems $1, 2, \ldots, t-1$ (in k to i comparisons, as in (3)) than it is for systems $t, t+1, \ldots, k$ to do so. The third inequality is a property of the procedure. □

If we know that the systems will be simulated independently (no CRN), then it is possible to reduce the value of η somewhat using an approach similar to Hartmann [1991]; all else being equal, the smaller η is the more quickly the procedure terminates. In this case, we set $g(\eta) = 1 - (1-\alpha)^{1/(k-1)}$ rather than $\alpha/(k-1)$, but otherwise leave the procedure unchanged. It is not difficult to show that $\alpha/(k-1) \le 1 - (1-\alpha)^{1/(k-1)}$ for all $0 < \alpha < 1$ and $k \ge 2$, and that $g^{-1}(\beta)$ is a nonincreasing function of β.

Theorem 2. *Under the same assumptions as Theorem* 1, *except that Σ is a diagonal matrix, the fully sequential indifference-zone procedure selects system k with probability $\ge 1 - \alpha$ when η solves $g(\eta) = 1 - (1-\alpha)^{1/(k-1)}$.*

PROOF. Let CS_i denote the event that a correct selection is made if system i faces system k in isolation. Then

$$\Pr\{\mathrm{CS}\} \geq \Pr\left\{\bigcap_{i=1}^{k-1} \mathrm{CS}_i\right\}$$

because the intersection event requires system k to eliminate each inferior system i individually, whereas in reality some system $\ell \neq i, k$ could eliminate i. Thus,

$$\begin{aligned}
\Pr\{\mathrm{CS}\} &\geq \Pr\left\{\bigcap_{i=1}^{k-1} \mathrm{CS}_i\right\} \\
&= \mathrm{E}\left[\Pr\left\{\bigcap_{i=1}^{k-1} \mathrm{CS}_i \,\middle|\, X_{k1}, \ldots, X_{k,N_k+1}, S_{1k}^2, \ldots, S_{k-1,k}^2\right\}\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{k-1} \Pr\left\{\mathrm{CS}_i \,\middle|\, X_{k1}, \ldots, X_{k,N_k+1}, S_{ik}^2\right\}\right], \qquad (6)
\end{aligned}$$

where the last equality follows because the events are conditionally independent. Clearly, (6) does not increase if we assume the slippage configuration, so we do so from here on.

Now notice that $\Pr\{\mathrm{CS}_i \mid X_{k1}, \ldots, X_{k,N_k+1}, S_{ik}^2\}$ is nondecreasing in X_{kj} and S_{ik}^2. Therefore, by Lemma 2,

$$\begin{aligned}
&\mathrm{E}\left[\prod_{i=1}^{k-1} \Pr\{\mathrm{CS}_i | X_{k1}, \ldots, X_{k,N_k+1}, S_{ik}^2\}\right] \\
&\geq \prod_{i=1}^{k-1} \mathrm{E}\left[\Pr_{\mathrm{SC}}\{\mathrm{CS}_i | X_{k1}, \ldots, X_{k,N_k+1}, S_{ik}^2\}\right] \\
&= \prod_{i=1}^{k-1} \mathrm{E}\left[1 - \Pr_{\mathrm{SC}}\{\mathrm{ICS}_i | S_{ik}^2\}\right] \\
&\geq \prod_{i=1}^{k-1} \left\{1 - \left(1 - (1-\alpha)^{1/(k-1)}\right)\right\} = 1 - \alpha
\end{aligned}$$

where the last inequality follows from the proof of Theorem 1. □

Remark 3. A corollary analogous to Corollary 1 can easily be proven in this case as well.

Key to the development of the fully sequential procedure is Fabian's [1974] result that allows us to control the chance that we incorrectly eliminate the best system, system k, when the partial sum process $\sum_{j=1}^{r}(X_{kj} - X_{ij})$ wanders too far below 0. Fabian's analysis is based on linking this partial sum process to a corresponding Brownian motion process. We are aware of at least one other large-deviation type result that could be used for this purpose [Robbins 1970], and in fact is used by Swanepoel and Geertsema [1976] to derive a fully sequential procedure. However, it is easy to show that Robbins' result leads to a continuation region with area that is, in expected value, much larger than Fabian's region; in fact, Robbins' region typically *contains* Fabian's region, suggesting that more data will be required to reach a conclusion with the Robbins' region.

3. DESIGN OF THE PROCEDURE

In this section, we examine factors that the experimenter can control in customizing the fully sequential procedure for their problem. Specifically, we look at the choice of c, whether or not to use CRN, and the effect of batch size. We conclude that $c = 1$ is a good compromise choice, CRN should almost always be employed, but the best batch size depends on the relative cost of stages of sampling versus individual observations.

3.1 Choice of c

Fabian's result defines a continuation region for the partial sum process, $\sum_{j=1}^{r}(X_{kj} - X_{ij})$. Provided $c < \infty$, this region is a triangle, and as long as the partial sum stays within this triangle sampling continues. As c increases the triangle becomes longer, but narrower, and in the limit becomes two parallel lines. Figure 1 shows the continuation region for our procedure.

The type of region that is best for a particular problem will depend on characteristics of the problem: If there is one dominant system and the others are grossly inferior, then having the region as narrow as possible is advantageous since the inferior systems will be eliminated quickly. However, if there are a number of very close competitors so that sampling is likely to continue to the end stage, then a short, fat region is desirable. Of course, the experimenter may not know such things in advance.

To compare various values of c, we propose looking at the *area* of the continuation region they imply. If we rotate the continuation region in Figure 1 ninety degrees counterclockwise, then the smallest area results from the best combination of small base—implying that clearly inferior systems can be eliminated early in the experiment—and short height—implying reasonable termination of the procedure if it goes to the last possible stage.

Using area as the metric immediately rules out $c = \infty$, since the area is infinite. For $c < \infty$, the area of the continuation region is (ignoring rounding)

$$\left(\frac{h^2 S_{ik}^2}{2c\delta}\right)\left(\frac{h^2 S_{ik}^2}{\delta^2}\right) = c\eta^2 \times \frac{2(n_0 - 1)^2 S_{ik}^4}{\delta^3}$$

or simply $c\eta^2$ in units of $2(n_0 - 1)^2 S_{ik}^4/\delta^3$. Thus, for fixed n_0, the key quantity is $c\eta^2$. Below, we provide numerical results that suggest that $c = 1$ minimizes this quantity over all k.

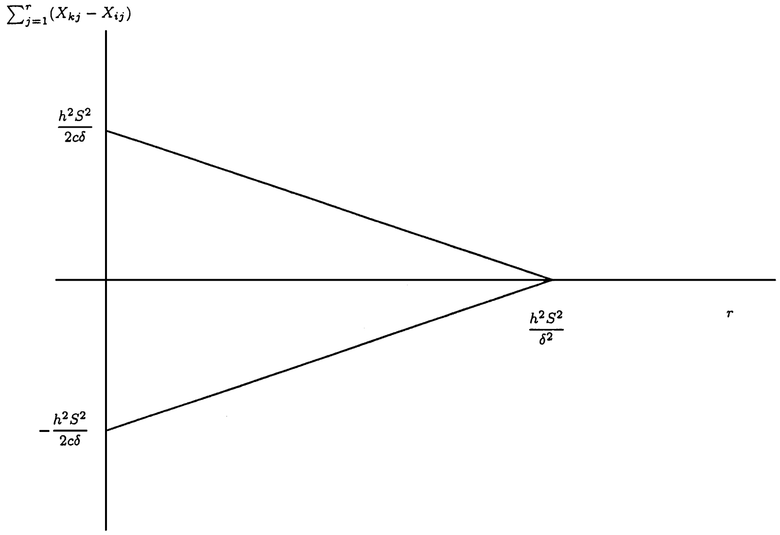
$\sum_{j=1}^{r}(X_{kj} - X_{ij})$
$\frac{h^2S^2}{2c\delta}$
$\frac{h^2S^2}{\delta^2}$
r
$-\frac{h^2S^2}{2c\delta}$

Fig. 1. Continuation region for the fully sequential, indifference-zone procedure when $c < \infty$.

Table I lists the value of $c\eta^2$ for $k = 2, 5, 10, 20, 100$ and $n_0 = 5, 10, 15, 20$, when $\beta = \alpha/(k-1)$ and $\alpha = 0.05$ (completely analogous results are obtained when $\beta = 1 - (1-\alpha)^{1/(k-1)}$). In all of these cases, $c = 1$ gives the smallest area. Therefore, when the experimenter has no idea if there are a few dominant systems or a number of close competitors, the choice of $c = 1$ appears to be a good compromise solution. We set $c = 1$ throughout the remainder of the paper.

3.2 Common Random Numbers

As described in Section 2, the choice to use or not to use CRN alters the fully sequential procedure only through the parameter η, which is smaller if we simulate the systems independently (no CRN). A smaller value of η tends to make the procedure more efficient. However, if we use CRN, then we expect the value of $S_{i\ell}^2$ to be smaller, which also tends to make the procedure more efficient. In this section, we show that even a small decrease in $S_{i\ell}^2$ due to the use of CRN is enough to offset the increase in η that we incur.

Recall that the parameter η is the solution of the equation $g(\eta) = \beta$ where $\beta = \alpha/(k-1)$ if we use CRN, while $\beta = 1-(1-\alpha)^{1/(k-1)}$ if we know the systems are simulated independently. Let η_C be the solution when CRN is employed, and let η_I be the solution if the systems are simulated independently.

For $c = 1$, it is easy to show that $g(\eta)$ is nonincreasing in $\eta > 0$; in fact, $g(\eta)$ is decreasing in $\eta > 0$ when $k \geq 3$. Further, $\alpha/(k-1)$ is always less than or equal to $1-(1-\alpha)^{1/(k-1)}$, with equality holding only when $k = 2$. These two facts imply that $\eta_C \geq \eta_I$. Below we derive a bound on η_C/η_I for a range of values of k, n_0 and α.

Table I. Area $c\eta^2$ in Units of $2(n_0-1)^2 S_{ik}^4\,/\delta^3$, when $g(\eta)=0.05/(k-1)$

				k		
n_0	c	2	5	10	20	100
5	1	1.169	7.088	18.007	40.858	232.017
	2	1.593	9.325	23.446	52.880	298.090
	3	2.096	12.140	30.433	68.517	385.426
	4	2.618	15.087	37.769	84.972	477.501
	5	3.147	18.084	45.237	101.701	571.359
	10	5.822	33.284	83.142	186.797	1048.228
10	1	0.112	0.403	0.738	1.221	3.296
	2	0.148	0.504	0.903	1.469	3.886
	3	0.193	0.645	1.148	1.863	4.893
	4	0.240	0.796	1.411	2.286	5.993
	5	0.288	0.946	1.682	2.723	7.116
	10	0.529	1.731	3.058	4.942	12.905
15	1	0.038	0.120	0.203	0.311	0.705
	2	0.050	0.148	0.243	0.366	0.804
	3	0.065	0.188	0.307	0.459	1.002
	4	0.080	0.231	0.376	0.563	1.222
	5	0.096	0.275	0.447	0.666	1.449
	10	0.176	0.499	0.810	1.204	2.611
20	1	0.019	0.056	0.092	0.136	0.285
	2	0.025	0.068	0.108	0.158	0.319
	3	0.032	0.087	0.136	0.197	0.395
	4	0.039	0.106	0.166	0.240	0.482
	5	0.047	0.126	0.198	0.286	0.571
	10	0.086	0.231	0.357	0.515	1.030

Ignoring rounding,

$$\begin{aligned}\Psi &\equiv \frac{\eta_C}{\eta_I}\\ &= \frac{(1/2)\left\{(2\alpha/(k-1))^{-2/(n_0-1)}-1\right\}}{(1/2)\left\{\left[2\left(1-(1-\alpha)^{1/(k-1)}\right)\right]^{-2/(n_0-1)}-1\right\}}\\ &= \frac{(2\alpha/(k-1))^{-2/(n_0-1)}-1}{\left[2\left(1-(1-\alpha)^{1/(k-1)}\right)\right]^{-2/(n_0-1)}-1}. \qquad (7)\end{aligned}$$

The ratio Ψ is a function of n_0, α and k, and we are interested in finding an upper bound for $10 \le n_0 \le 20$, $0 < \alpha \le 0.1$ and $2 \le k \le 100$, the range of parameters we consider to be of practical importance.

To accomplish this, we evaluated the $\partial\Psi/\partial n_0$ for all $k = 2, 3, \ldots, 100$, $n_0 = 2, 3, \ldots, 20$, and on a narrow grid of α values (including the standard 0.10, 0.05 and 0.01 values). For this range, $\partial\Psi/\partial n_0$ is always less than zero; therefore, Ψ seems to be a decreasing function of $n_0 \ge 2$. This implies that we need to consider only the smallest value of n_0 to find an upper bound on (7).

After setting $n_0 = 10$, the smallest value of interest to us, we observed that (7) is an increasing function of α by evaluating $\partial\Psi/\partial\alpha$ for each k in the range of interest. Thus, the largest α, which is 0.1, should be chosen to find an upper bound:

$$\Psi \overset{\substack{n_0=10,\\ \alpha=0.1}}{\le} \frac{-1+1.42997(1/(k-1))^{-2/9}}{-1+\left\{2\left(1-0.9^{1/(k-1)}\right)\right\}^{-2/9}}. \tag{8}$$

Now we have only one variable remaining, k, and by evaluating (8) for all k in the range of interest we find that $k=7$ gives the largest value, which is 1.01845. Therefore, we can say that for our range of interest

$$1 \le \frac{\eta_C}{\eta_I} < 1.02.$$

That is, η_C is at most 1.02 times η_I. To relate this ratio to the potential benefits of using CRN, we consider two performance measures: the expected maximum number of observations and the expected area of the continuation region.

Let N_C and N_I be the maximum number of replications until the procedure terminates, and let A_C and A_I be the area of the continuation region with CRN and without CRN, respectively. To simplify the presentation, let $k=2$, assume the variances across systems are all equal to σ^2 and that the correlation induced between systems by CRN is $\mathrm{Corr}[X_{ij}, X_{\ell j}] = \rho > 0$ (taking $k > 2$ does not change the result). Then, ignoring rounding,

$$\frac{E[N_C]}{E[N_I]} = \frac{2c\eta_C(n_0-1)2\sigma^2(1-\rho)/\delta^2}{2c\eta_I(n_0-1)2\sigma^2/\delta^2} = \frac{\eta_C}{\eta_I}(1-\rho). \tag{9}$$

Equation (9) shows that CRN will reduce the expected maximum number of replications if

$$\rho > 1 - \frac{\eta_I}{\eta_C} \tag{10}$$

implying that $\rho \ge 1 - 1/1.02 \doteq 0.02$ is always sufficient on our range of interest.

Recall that the area of the continuation region for a pair of systems i, j is given by $2c\eta^2(n_0-1)^2 S_{ij}^4/\delta^3$. Under the same assumptions, we can show that

$$\frac{E[A_C]}{E[A_I]} = \frac{\eta_C^2}{\eta_I^2}(1-\rho)^2. \tag{11}$$

This again implies that $\rho \ge 1 - 1/1.02 \doteq 0.02$ is sufficient, on our range of interest, for CRN to reduce the expected area of the continuation region. Therefore, for the range of parameters $2 \le k \le 100$, $10 \le n_0 \le 20$, and $0.01 \le \alpha \le 0.10$, we claim that achieving a positive correlation of at least 0.02 is sufficient to make the use of CRN superior to simulating the systems independently.

3.3 The Effect of Batch Size

There are several reasons why an experimenter might want the jth observation from system i to be the mean of a batch of more basic observations:

- When the computing overhead for switching among simulated systems is high, it may be computationally efficient to obtain more than one observation from each system each time it is simulated.
- Even if the basic observations deviate significantly from the assumed normal distribution, batch means of these observations will typically be more nearly normal.

—If the simulation of system i involves only a single, incrementally extended, replication of a steady-state simulation, then the basic observations may deviate significantly from the assumed independence, while batch means of a sufficiently large number of basic observations may be nearly independent.

In this section, we investigate the effect of batch size on the fully sequential procedure. To facilitate the analysis, we assume that the total number of basic observations obtained in the first-stage of sampling, denoted n_0^{raw}, is fixed, and that all systems use a common batch size b. However, the procedure itself does not depend on a common batch size across systems, only that the batch size remains fixed within each system.

Let X_{ij} denote a basic observation, and $X_{ij}[b]$ denote the mean of a batch of b basic observations; we use the batch means $X_{ij}[b]$ instead of the basic observations in the Fully Sequential Indifference-Zone Procedure of Section 2. Let $S_{i\ell}^2[b]$ be the sample variance of the difference between the batch means from systems i and ℓ, and let $n_0 = n_0^{\text{raw}}/b$ denote the number of batch means $X_{ij}[b]$ used in the first stage of sampling. For the purpose of analysis, we assume that n_0 is always integer and that for each system i, the basic observations X_{ij}, $j = 1, 2, \ldots$, are independent and follow a normal distribution with mean μ_i and common unknown variance σ^2. We make the assumption of equal variances to simplify the presentation; recall that equal variances is not an assumption of our procedure, and the results below do not change—except for the units on them—if the variances are unequal.

Under these conditions,

$$\sigma_{i\ell}^2[b] \equiv \text{Var}[X_{ij}[b] - X_{\ell j}[b]] = \frac{2\sigma^2}{b} = \frac{2n_0\sigma^2}{n_0^{\text{raw}}}$$

and

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

If we ignore rounding,

$$\begin{aligned} \text{E}[N_i] &= \frac{2\eta \times (n_0 - 1)}{\delta^2}\text{E}\Big[\max_{\ell \neq i} S_{i\ell}^2[b]\Big] \\ &= 4\eta n_0 \text{E}\left[\max_{\ell \neq i}\frac{(n_0 - 1)S_{i\ell}^2[b]}{\sigma_{i\ell}^2[b]}\right] \times \left(\frac{\sigma/\sqrt{n_0^{\text{raw}}}}{\delta}\right)^2. \end{aligned} \tag{12}$$

As mentioned in Section 2, $(n_0 - 1)S_{i\ell}^2[b]/\sigma_{i\ell}^2[b]$ has a chi-squared distribution with $n_0 - 1$ degrees of freedom, so the expected maximum number of stages involves the maximum of $k - 1$ identically distributed, but dependent, chi-squared variables. However, simulation analysis of several cases revealed that the effect of this dependence on the expected value is weak, so for the purpose of analysis we use the approximation

$$\text{E}\left[\max_{\ell \neq i}\frac{(n_0 - 1)S_{i\ell}^2[b]}{\sigma_{i\ell}^2[b]}\right] \approx \int_0^\infty y(k-1)\{F(y)\}^{k-2}f(y)dy, \tag{13}$$

Table II. $\mathrm{E}[N_i]$, in Units of $(\sigma/(\delta\sqrt{n_0^{\mathrm{raw}}}))^2$, as a Function of Number of Batches n_0 when $c = 1$, $1-\alpha = 0.95$ and the Number of Observations n_0^{raw} is Fixed

	k				
n_0	2	5	10	20	100
2	396	15206	112590	664751	28653750
3	108	956	2953	7853	59910
4	87	479	1111	2271	9714
5	86	373	754	1356	4376
6	91	339	631	1053	2882
7	97	328	579	9207	2262
8	104	329	557	855	1950
9	112	334	550	823	1775
10	120	343	552	809	1671
11	129	354	559	806	1608
12	137	366	569	810	1572
13	146	379	582	819	1552
14	155	393	596	831	1543
15	164	408	612	846	1544
16	172	422	629	862	1550
17	181	437	646	880	1562
18	190	453	664	899	1577
19	199	468	682	919	1594
20	208	483	701	940	1615
21	217	499	719	961	1637
22	227	515	739	982	1660
23	236	530	758	1004	1685
24	245	546	777	1027	1711

where f and F are the density and cdf, respectively, of the chi-square distribution with $n_0 - 1$ degrees of freedom. In other words, we treat the $S_{i\ell}^2[b]$, $\ell \neq i$ as if they are independent.

Expression (13) is a function of k and n_0, while η is a function of k, n_0 and α. If we assume that n_0^{raw} is given, then for any fixed k and $1-\alpha$ the expected maximum number of stages (12) depends only on the initial number of batches, n_0, provided we express it in units of $(\sigma/(\delta\sqrt{n_0^{\mathrm{raw}}}))^2$. Unfortunately, there is no closed-form expression for (13), but we can evaluate it numerically.

Table II gives $\mathrm{E}[N_i]$ as a function of the number of batches n_0 for different values of k. This table shows that the expected maximum number of *stages* decreases, then increases, as a function of the number of batches in the first stage when n_0^{raw} is fixed. The savings in the beginning are caused by increasing the degrees of freedom. However, after n_0 passes some point there is no further benefit from increased degrees of freedom, so the effect of increasing n_0 (dividing the output into smaller batches, eventually leading to a batch size of 1) is simply to increase the number of stages. The expected maximum number of stages, $\mathrm{E}[N_i]$, is important when the computing overhead for switching among systems is substantial.

Table III shows $\mathrm{E}[bN_i]$, the number of basic (unbatched) observations, as the number of batches (but not the number of basic first-stage observations n_0^{raw}) is increased. The expected maximum number of basic observations is important when the cost of obtaining basic observations dominates the cost of multiple

Table III. E[bN_i], in Units of $(\sigma/(\delta\sqrt{n_0^{\text{raw}}}))^2$, as a Function of Number of Batches n_0 when $c = 1$, $1 - \alpha = 0.95$ and the Number of Observations n_0^{raw} is Fixed

			k		
n_0	2	5	10	20	100
2	4752	182471	1351084	7977012	343844999
3	864	7648	23622	62824	479281
4	524	2874	6666	13624	58284
5	415	1793	3619	6511	21005
6	363	1356	2522	4212	11526
7	332	1126	1984	3155	7757
8	313	986	1670	2565	5851
9	299	891	1467	2195	4733
10	289	823	1324	1942	4011
11	281	772	1219	1758	3509
12	275	733	1138	1620	3143
13	269	701	1074	1512	2864
14	265	674	1022	1425	2646
15	262	652	979	1353	2470
16	259	634	943	1293	2326
17	256	617	912	1243	2205
18	254	603	885	1199	2102
19	252	591	861	1161	2014
20	250	580	841	1128	1938
21	249	570	822	1098	1870
22	247	561	806	1072	1811
23	246	553	791	1048	1758
24	245	546	777	1027	1711

stages of sampling. As the table shows, this quantity is minimized by using a batch size of $b = 1$ (i.e., no batching), but the table also shows that once we obtain, say, 15 to 20 batches, there is little potential reduction in E[bN_i] from increasing the number of batches (decreasing the batch size) further.

For a fixed number of basic observations in the first stage, the choice of number of batches n_0 should be made considering the two criteria, E[N_i] and E[bN_i]. If the cost of switching among systems dominates, then a small number of batches (5 to 10) will tend to minimize the number of stages. When the cost of obtaining each basic observation dominates, as it often will, then from 15 to 20 batch means are desirable at the first stage; of course, if neither nonnormality nor dependence is a problem then a batch size of 1 will be best in this case. Damerdji and Nakayama [1999] address the impact of batch size when the number of basic observations can increase as a function of δ.

4. EXPERIMENTS

In this section, we summarize the results of experiments performed to compare the following procedures:

(1) Rinott's [1978] procedure (RP), a two-stage indifference-zone selection procedure that makes no attempt to eliminate systems prior to the second (and last) stage of sampling.

(2) A two-stage screen-and-select procedure (2SP) proposed by Nelson et al. [2001] that uses subset selection (at confidence level $1 - \alpha/2$) to eliminate systems after the first stage of sampling, and then applies Rinott's second-stage sampling rule (at confidence level $1 - \alpha/2$) to the survivors.

(3) The fully sequential procedure (FSP) proposed in Section 2, both with and without CRN (recall that the two versions differ only in the value of η used).

The systems were represented by various configurations of k normal distributions; in all cases, system 1 was the true best (had the largest true mean). We evaluated each procedure on different variations of the systems, examining factors including the number of systems, k; batch size, b; the correlation between systems, ρ; the true means, $\mu_1, \mu_2, \ldots, \mu_k$; and the true variances, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. The configurations, the experiment design, and the results are described below.

4.1 Configurations and Experiment Design

To allow for several different batch sizes, we chose the first-stage sample size to be $n_0^{\mathrm{raw}} = 24$, making batch sizes of $b = 1$, 2 or 3 possible. Thus, n_0 (the number of first-stage batch means) was 24, 12 or 8, respectively. The number of systems in each experiment varied over $k = 2, 5, 10, 25, 100$.

The indifference zone, δ, was set to $\delta = \sigma_1/\sqrt{n_0^{\mathrm{raw}}}$, where σ_1^2 is the variance of an observation from the best system. Thus, δ is 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 $\mu_2 = \mu_3 = \cdots = \mu_k = 0$. This is a difficult configuration for procedures that try to eliminate systems because all of the inferior systems are close to the best. To investigate the effectiveness of the procedures in eliminating noncompetitive systems, monotone decreasing means (MDM) were also used. In the MDM configuration, the means of all systems were spaced evenly apart according to the following formula: $\mu_i = \mu_1 - a(i-1)$, for $i = 2, 3, \ldots, k$, where $a = \delta/\tau$. Values of τ were $\tau = 1/2$, 1 or 3 (effectively spacing each mean 2δ, δ or $\delta/3$ from the previous mean).

For each configuration of the means, we examined the effect of both equal and unequal variances. In the equal-variance configuration, σ_i was set to 1. In the unequal-variance configuration, the variance of the best system was set both higher and lower than the variances of the other systems. In the MDM configurations, experiments were run with the variance directly proportional to the mean of each system, and inversely proportional to the mean of each system. Specifically, $\sigma_i^2 = |\mu_i - \delta| + 1$ to examine the effect of increasing variance as the mean decreases, and $\sigma_i^2 = 1/(|\mu_i - \delta| + 1)$ to examine the effect of decreasing variances as the mean decreases. In addition, some experiments were run with means in the SC, but with the variances of all systems either monotonically decreasing or monotonically increasing as in the MDM configuration.

When CRN was employed we assumed that the correlation between all pairs of systems was ρ, and values of $\rho = 0.02, 0.25, 0.5, 0.75$ were tested. Recall that $\rho = 0.02$ is the lower bound on correlation that we determined is necessary to insure that the FSP with CRN is at least as efficient as the FSP without CRN.

Thus, we had a total of six configurations: SC with equal variances, MDM with equal variances, MDM with increasing variances, MDM with decreasing variances, SC with increasing variances and SC with decreasing variances. For each configuration, 500 macroreplications (complete repetitions) of the entire experiment were performed. In all experiments, the nominal probability of correct selection was set at $1 - \alpha = 0.95$. To compare the performance of the procedures we recorded the total number of basic (unbatched) observations required by each procedure, and the total number of stages, on each macroreplication and reported the sample averages (when data are normally distributed all of the procedures achieve the nominal probability of a correct selection).

4.2 Summary of Results

The overall experiments showed that the FSP is superior to the other procedures across all of the configurations we examined. Under difficult configurations, such as SC with increasing variances, FSP's superiority relative to RP and 2SP was more noticeable as the number of systems increased.

As we saw in Table III the expected maximum number of basic observations that might be taken from system i, $\mathrm{E}[bN_i]$, increases as batch size increases (number of batches decreases); this was borne out in the experiments as the total *actual* number of observations taken also increased as batch size b increased. However, the total number of basic observations increased more slowly for the FSP than for RP or 2SP as batch size increased. The number of stages behaved as anticipated from our analysis of the expected maximum number of stages: first decreasing, then increasing as the number of batches increases.

Finally, and not unexpectedly, in the MDM configuration wider spacing between the true means made both 2SP and FSP work better (eliminate more systems earlier) than they did otherwise.

4.3 Some Specific Results

Instead of presenting comprehensive results from such a large simulation study, we present selected results that emphasize the key conclusions.

4.3.1 *Effect of Number of Systems.* In our experiments, the FSP outperformed all of the other procedures under every configuration; see Tables IV and V for illustrations. Reductions of more than 50% in the number of basic observations, as compared to RP and 2SP, were obtained in most cases. As the number of systems increased under difficult configurations—such as MDM or SC with increasing variances—the benefit of FSP relative to RP and 2SP was even greater.

4.3.2 *Effect of Batch Size.* Results in Section 3.3 suggest that the total number of basic observations should be an increasing function of the batch size b (a decreasing function of the number of batches n_0^{raw}/b), while the number of stages should decrease, then increase in b. Tables IV and V show empirical results for the total number of basic observations, while Table VI shows the total number of stages, for different values of b. As expected, the total number of basic observations for the FSP (as well as for RP and 2SP) is always increasing

Table IV. Sample Average of the Total Number of Basic (Unbatched) Observations when the Number of Systems is $k = 5$ and the Spacing between the Means is $\delta/\tau = \delta$ as a Function of Batch Size b and Induced Correlation (ρ)

	MDM			MDM			SC			SC		
Procedure	increasing var			decreasing var			increasing var			decreasing var		
b	1	2	3	1	2	3	1	2	3	1	2	3
RP	1894	2217	2770	998	1168	1455	1894	2217	2770	998	1168	1455
2SP	1620	2179	3138	794	1022	1477	2187	2714	3643	1101	1381	1871
FSP(0)	542	659	808	403	481	580	939	1126	1414	594	721	888
FSP(0.02)	530	665	808	387	481	583	939	1141	1388	569	706	867
FSP(0.25)	416	498	625	308	368	442	731	867	1106	449	536	668
FSP(0.50)	289	349	429	220	252	319	485	607	741	303	363	442
FSP(0.75)	166	193	240	147	164	192	255	311	386	172	205	247

Table V. Sample Average of the Total Number of Basic (Unbatched) Observations when the Number of Systems is $k = 25$ and the Spacing between the Means is $\delta/\tau = \delta$ as a Function of Batch Size b and Induced Correlation (ρ)

	MDM			MDM			SC			SC		
Procedure	increasing var			decreasing var			increasing var			decreasing var		
b	1	2	3	1	2	3	1	2	3	1	2	3
RP	39505	49207	66962	4272	5325	7237	39506	49207	66962	4272	5325	7237
2SP	4644	8514	18957	1550	2210	3533	45364	59190	85536	4185	5892	8911
FSP(0)	2009	2542	3413	1157	1390	1784	13399	17710	23601	2312	2887	3830
FSP(0.02)	1965	2498	3389	1143	1368	1756	13241	17446	23145	2304	2872	3753
FSP(0.25)	1554	1971	2625	983	1153	1439	10736	13771	18562	1792	2259	3033
FSP(0.50)	1118	1412	1870	813	923	1103	7750	10145	13690	1288	1621	2154
FSP(0.75)	749	900	1152	667	717	806	4960	6180	8192	822	996	1264

Table VI. Sample Average of the Total Number of Stages when Number of Systems is $k = 25$ and Spacing between the Means is $\delta/\tau = \delta$ as a Function of Batch Size b and Induced Correlation (ρ)

	MDM			MDM			SC			SC		
Procedure	increasing var			decreasing var			increasing var			decreasing var		
b	1	2	3	1	2	3	1	2	3	1	2	3
FSP(0)	240	171	173	189	136	136	1229	927	921	229	163	171
FSP(0.02)	234	167	173	184	130	131	1197	898	895	228	168	164
FSP(0.25)	173	122	124	137	98	99	971	707	715	167	126	131
FSP(0.50)	107	79	82	83	61	63	708	535	558	110	85	89
FSP(0.75)	45	37	41	31	26	28	475	342	350	51	43	45

in b, with the incremental increase becoming larger as b increases. On the other hand, the number of stages is usually minimized at $b = 2$, as shown in Table VI. The number of stages is always 1 or 2 for RP and 2SP.

The total number of basic observations taken by RP and 2SP is more sensitive to the batch size than FSP; see Table V, for example. Under the MDM with increasing variances, the total number of basic observations taken by 2SP when $b = 3$ was more than four times larger than when $b = 1$. However, for the FSP, the number of basic observations was only 1.5 times larger when moving from $b = 1$

Table VII. Sample Average of the Total Number of Basic (Unbatched) Observations for the FSP when the Number of Systems is $k = 25$, Spacing between the Means is $\delta/\tau = 2\delta$ and the Batch Size is $b = 1$ as a Function of Correlation ρ

	MDM	MDM	SC	SC
ρ	increasing var	decreasing var	increasing var	decreasing var
0	1504	828	21778	1725
0.02	1490	827	21090	1689
0.25	1200	755	17543	1367
0.50	919	681	13801	1037
0.75	675	617	9067	762

Table VIII. Sample Average of the Total Number of Basic (Unbatched) Observations for the FSP when the Number of Systems is $k = 10$, the Batch Size is $b = 1$ and the Systems are Simulated Independently as a Function of the Spacing between the Means δ/τ

	MDM	MDM	SC	SC
δ/τ	increasing var	decreasing var	increasing var	decreasing var
2δ	681	400	3945	943
δ	981	630	2868	1149
$\delta/3$	1657	1311	2126	1452

to $b = 3$. This effect becomes even more pronounced as the number of systems becomes larger.

4.3.3 *Effect of Correlation.* Results in Section 3.2 suggest that positive correlation larger than 0.02 is sufficient for the FSP with CRN to outperform the FSP assuming independence. As shown in the empirical results in Table VII, FSP under independence is essentially equivalent to the FSP under CRN when $\rho = 0.02$ in terms of the number of basic observations. Of course, a larger positive correlation makes the FSP even more efficient, and this holds across all of the configurations that were used in our experiments.

4.3.4 *Effect of Spacing.* In our experiments, spacing between means was defined by multiples of δ/τ, so that small τ implies large spacing. Larger spacing makes it easier for any procedure that screens or eliminates systems to remove inferior systems. Table VIII shows that, in most cases, the total number of basic observations for the FSP decreases as τ/δ increases. The exception is the SC with increasing variances where the FSP actually does worse with wider spacing of the means (this happened for all values of k and b, and for 2SP as well as FSP). A similar pattern emerged for the total number of stages.

To explain the counterintuitive results for the SC with increasing variances, recall that in this configuration all inferior systems have the same true mean, but the variances are assigned as in the MDM configuration with increasing variances; that is, the variance of the ith system is $\sigma_i^2 = |\mu_i - \delta| + 1$ with μ_i as it *would be* in the MDM configuration. Therefore, larger δ/τ implies larger spacing, and larger spacing implies variances that increase much faster. Thus, in this example, the effect of increasing the variances of the inferior systems is greater than the effect of spacing the means farther apart. This is consistent with what we have seen in other studies we have conducted: inferior systems with large variances provide difficult cases for elimination procedures.

4.4 Robustness Study

In practical computer simulation experiments, the FSP may be applied when the simulation output data from each system are neither normally distributed nor independent (the procedure accounts for dependence across systems due to CRN). We have recommended batching as a way to mitigate these conditions, and analyzed the effect of batch size on the FSP. To more directly assess the impact of departures from normality and independence, we also performed a small-scale robustness study in which the FSP was applied to dependent, exponentially distributed data. Specifically, we let the output data from system i be defined by the EAR(1) process

$$X_{ij} = \begin{cases} \phi X_{i,j-1}, & \text{with probability } \phi \\ \phi X_{i,j-1} + \varepsilon_{ij}, & \text{with probability } 1-\phi \end{cases} \tag{14}$$

where the ε_{ij} are independent and identically exponentially distributed with mean $1/\lambda_i$ and $0 \leq \phi < 1$ (see Lewis [1980]). CRN was not employed in these experiments. For the EAR(1) process, $\mathrm{Corr}[X_{ij}, X_{i,j+h}] = \phi^h$, and we used a common value of $\phi = 0, 0.5$ or 0.9 for all systems. The model produces data that are marginally exponential with mean $1/\lambda_i$, and when $\phi = 0$ the data are independent and identically distributed.

We defined the alternative systems by setting different values for λ_i, which changes both the mean and the variance of the output data. Here, we report results for two versions of the slippage configuration:

SCMAX: In this configuration, we set $1/\lambda_1 = 1$ and $1/\lambda_2 = \cdots = 1/\lambda_k = 1 - \delta$. Thus, a larger mean is better, and the best system has the largest variance.

SCMIN: In this configuration, we set $1/\lambda_1 = 1$ and $1/\lambda_2 = \cdots = 1/\lambda_k = 1 + \delta$. Thus, a smaller mean is better, and the best system has the smallest variance.

We set $\delta = \sigma_\infty^2/\sqrt{n_0^{\mathrm{raw}}}$, where $\sigma_\infty^2 = \lim_{r\to\infty} r\mathrm{Var}[\bar{X}_1(r)]$ is the asymptotic variance of the best system, system 1. When $\phi = 0$, this coincides with our definition of δ in the normal distribution experiments.

In the experiments reported here, we fixed the number of systems at $k = 5$, the confidence level at $1 - \alpha = 0.95$, and the number of first-stage batch means at $n_0 = 10$. We let n_0^{raw} vary from 10 (which implies batch size $b = 1$, or no batching) to 1000 (implying $b = 100$). We made no attempt to use the output data to determine an appropriate batch size; this is a topic of current research.

Overall, we found dependence to be a more serious problem than nonnormality. When there is only weak (or no) dependence, the FSP attained near the desired PCS with little or no batching. However, a large batch size was required to overcome the effect of strong positive autocorrelation.

Tables IX and X report the sample average of the total number of basic (unbatched) observations ("RawRep") and the total number of stages ("Stage"), as well as the estimated PCS, for different batch sizes b and levels of dependence ϕ. Remember that in these experiments $n_0^{\mathrm{raw}} = 10b$, but δ is adjusted to make them comparable. All results are based on 500 macroreplications. Notice that the SCMIN configuration demanded more raw replications and stages than

Table IX. Average Performance of the FSP for SCMAX when Data are EAR(1) With $1/\lambda_1 = 1$ and $1/\lambda_2 = \cdots = 1/\lambda_5 = 1 - \delta$

ϕ	b	RawRep	Stage	PCS
0	1	279	47	0.954
	2	600	51	1.000
	10	3570	62	1.000
	100	38301	68	1.000
0.5	1	72	5	0.772
	2	233	14	0.818
	3	476	23	0.854
	4	766	29	0.896
	5	1069	34	1.000
	10	2797	47	1.000
	100	36881	65	1.000
0.9	4	206	1	0.738
	7	476	5	0.760
	10	880	9	0.774
	100	30650	52	0.948

Table X. Average Performance of the FSP for SCMIN when the Data are EAR(1) With $1/\lambda_1 = 1$ and $1/\lambda_2 = \cdots = 1/\lambda_5 = 1 + \delta$

ϕ	b	RawRep	Stage	PCS
0	1	561	103	0.916
	2	1013	92	1.000
	10	4432	80	0.951
	100	40923	73	0.960
0.5	1	188	29	0.590
	2	537	45	0.736
	3	960	55	0.808
	4	1393	61	1.000
	5	1798	63	1.000
	10	4116	73	1.000
	100	41633	74	1.000
0.9	1	52	1	0.372
	10	2037	32	0.652
	20	6162	53	0.842
	100	40802	73	0.940

SCMAX, demonstrating again that difficult cases arise when the inferior systems have larger variances.

When $\phi = 0$, the data are exponentially distributed, but independent both within and across systems. Even with no batching (batch size $b = 1$) the estimated PCS is close to the nominal. When $\phi = 0.9$, so that the dependence is strong, the estimated PCS is well below the nominal level unless $b = 100$. In the presence of strong positive autocorrelation, the sample variance underestimates the true variance of the sample mean, leading to a continuation region that is too narrow. This causes the best system to be incorrectly eliminated at an early stage. As the batch size increases, the variance estimate becomes more accurate, the number of stages increases, and the PCS attains the desired level.

5. CONCLUSIONS

In this article, we presented a fully sequential, indifference-zone selection procedure that allows for unequal variances, batching and common random numbers. As we discussed in Section 4, the procedure is uniformly superior to two existing procedures across all the scenarios we examined, and it is significantly more efficient when the number of systems is large or the correlation induced via CRN is large. One advantage of the FSP is that it is easy to account for the effect of CRN, which is not true of 2SP, for instance (see Nelson et al. [2001] for a discussion of this point).

The results in this article suggest several possibilities for improving the FSP. One is to search for a tighter continuation region than the triangular one suggested by Fabian's lemma. A tighter region would seem to be possible since our estimates of the true probability of correct selection for the FSP (not reported here) show that it is typically greater than the nominal $1 - \alpha$.

Although we did consider the effect of batching, our results are most relevant for the situation in which we batch to reduce the number of stages or to improve the approximation of normality, rather than to mitigate the dependence in a single replication of a steady-state simulation. A small-scale robustness study suggests that the FSP may be somewhat tolerant of nonnormality, but not of dependence. We address this problem with procedures specifically designed to account for dependence within a single replication in Goldsman et al. [2002].

ACKNOWLEDGMENTS

The authors acknowledge the helpful comments and suggestions of the Associate Editor and referees.

REFERENCES

BECHHOFER, R. E., DUNNETT, C. W., GOLDSMAN, D. M., AND HARTMANN, M. 1990. A comparison of the performances of procedures for selecting the normal population having the largest mean when the populations have a common unknown variance. *Commun. Stat. B19*, 971–1006.

BECHHOFER, R. E., SANTNER, T. J., AND GOLDSMAN, D. M. 1995. *Design and Analysis for Statistical Selection*, *Screening and Multiple Comparisons*. Wiley, New York.

BOESEL, J., NELSON, B. L., AND KIM, S. 2001. Using ranking and selection to clean up after a simulation search. Tech. Rep. Department of Industrial Engineering and Management Sciences, Northwestern Univ., Evanston, Ill.

CHEN, C.-H. 1996. A lower bound for the correct-selection probability and its application to discrete event simulations. *IEEE Trans. Autom. Contr. 41*, 1227–1231.

CHICK, S. E., AND INOUE, K. 2001a. New procedures to select the best simulated system using common random numbers. *Manage. Sci. 47*, 1133–1149.

CHICK, S. E., AND INOUE, K. 2001b. New two-stage and sequential procedures for selecting the best simulated system. *Oper. Res. 49*, 732–743.

DAMERDJI, H., AND NAKAYAMA, M. K. 1999. Two-stage multiple-comparison procedures for steady-state simulations. *ACM Trans, Mod. Comput. Sim. 9*, 1–30.

FABIAN, V. 1974. Note on Anderson's sequential procedures with triangular boundary. *Ann. Statis. 2*, 170–176.

GOLDSMAN, D. M., KIM, S., MARSHALL, W., AND NELSON, B. L. 2002. Ranking and selection for steady-state simulation: Procedures and analysis. *INFORMS Journal on Computing*, Forthcoming.

GOLDSMAN, D. M., AND NELSON, B. L. 1998a. Comparing systems via simulation. In *Handbook of Simulation*, J. Banks, Ed., Chap. 8. Wiley, New York.

GOLDSMAN, D. M., AND NELSON, B. L. 1998b. Statistical screening, selection and multiple comparison procedures. In *Proceedings of the 1998 Winter Simulation Conference.* IEEE, Piscataway, N.J., pp. 159–166.

HARTMANN, M. 1988. An improvement on Paulson's sequential ranking procedure. *Sequen. Analysis 7*, 363–372.

HARTMANN, M. 1991. An improvement on Paulson's procedure for selecting the population with the largest mean from k normal populations with a common unknown variance. *Sequent. Analysis 10*, 1–16.

HOCHBERG, Y., AND TAMHANE, A. C. 1987. *Multiple Comparison Procedures*. J Wiley, New York.

HSU, J. C. 1996. *Multiple Comparisons: Theory and Methods*. Chapman and Hall, New York.

JENNISON, C., JOHNSTONE, I. M., AND TURNBULL, B. W. 1982. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III*, Vol 2. Academic Press, New York.

LEWIS, P. A. W. 1980. Simple models for positive-valued and discrete-valued time series with ARMA correlation structure. In *Multivariate Analysis V*, P. R. Krishnaiah, Ed. North-Holland, New York, pp. 151–156.

MILLER, J. O., NELSON, B. L., AND REILLY, C. H. 1996. Getting more from the data in a multinomial selection problem. In *Proceedings of the 1996 Winter Simulation Conference* J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain Eds. IEEE, Piscataway, N.J., pp. 287–294.

MILLER, J. O., NELSON, B. L., AND REILLY, C. H. 1998a. Efficient multinomial selection in simulation. *Naval Res. Log. 45*, 459–482.

MILLER, J. O., NELSON, B. L., AND REILLY, C. H. 1998b. Comparing simulated systems based on the probability of being the best. Tech. Rep., Dept. Industrial Engineering and Management Sciences, Northwestern Univ., Evanston, Ill.

NELSON, B. L., AND BANERJEE, S. 1999. Evaluating the probability of a good selection. In *Proceedings of the 1999 Winter Simulation Conference* P. A. Farrington, H. B. Nembhard, D. Sturrock, and G. W. Evans, Eds. IEEE, Piscataway, N.J., pp. 611–617.

NELSON, B. L., AND BANERJEE, S. 2001. Selecting a good system: Procedures and inference. *IIE Trans. 33*, 149–166.

NELSON, B. L., AND GOLDSMAN, D. M. 2001. Comparisons with a standard in simulation experiments. *Manage. Sci. 47*, 449–463.

NELSON, B. L., SWANN, J., GOLDSMAN, D. M., AND SONG, W. 2001. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Oper. Res.*, to appear.

PAULSON, E. 1964. A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Stat. 35*, 174–180.

RINOTT, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Commun. Stat. A7*, 799–811.

ROBBINS, H. 1970. Statistical methods related to the law of the iterated logarithm. *Ann. Math. Stat. 41*, 1397–1409.

SWANEPOEL, J. W. H., AND GEERTSEMA, J. C. 1976. Sequential procedures with elimination for selecting the best of k normal populations. *S. Afr. Statist. J. 10*, 9–36.

TAMHANE, A. C. 1977. Multiple comparisons in model I: One-way anova with unequal variances. *Commun. Stat. A6*, 15–32.

Received July 1999; revised June 2000; accepted July 2000