Comparisons with a Standard in Simulation Experiments

Barry L. Nelson • David Goldsman
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

We consider the problem of comparing a finite number of stochastic systems with respect to a single system (designated as the "standard") via simulation experiments. The comparison is based on expected performance, and the goal is to determine if any system has larger expected performance than the standard, and if so to identify the best of the alternatives. In this paper we provide two-stage experiment design and analysis procedures to solve the problem for a variety of scenarios, including those in which we encounter unequal variances across systems, as well as those in which we use the variance reduction technique of common random numbers *and it is appropriate to do so.* The emphasis is added because in some cases common random numbers can be counterproductive when performing comparisons with a standard. We also provide methods for estimating the critical constants required by our procedures, present a portion of an extensive empirical study, and demonstrate one of the procedures via a numerical example.
(*Simulation; Multiple Comparisons; Ranking and Selection; Output Analysis*)

1. Introduction

We consider an important case of the general class of problems that require comparing a finite and relatively small number of simulated systems in terms of their expected performance. By small we mean less than 20 systems.

There has been recent interest in the specific problem of determining the best of these systems, where "best" means maximum or minimum expected performance of a single common performance measure. See, for instance, Goldsman et al. (1991), Matejcik and Nelson (1995), Nakayama (1995), Nelson and Matejcik (1995), Yang and Nelson (1991), and the references therein. This work has yielded a rich collection of procedures, including two-stage procedures that guarantee with high probability (or confidence) that the best system is chosen, and simultaneously guarantee that confidence intervals for the difference between each alternative's performance and the best system's performance contain the true differences. In this paper we derive procedures with very similar characteristics that apply to problems in which one of the systems is singled out as the standard or benchmark, and the others are *evaluated with respect to the standard as well as with respect to each other*.

A basic rule of thumb when comparing systems is that sharper inference is obtained by focusing only on the specific comparisons that are relevant to the application at hand (Hsu 1996, Ch. 2). For instance, when it is important to select the best, and the number of observations is fixed, then the procedures cited above are more statistically efficient—meaning more likely to detect actual differences between each system and the best—than procedures that provide all pairwise comparisons among the alternatives. Similarly, when we desire comparisons with respect to a standard, we are more likely to obtain conclusive results if

0025-1909/01/4703/0449$5.00
1526-5501 electronic ISSN

we derive procedures that specifically deliver those comparisons.

In many applications the expected performance of the standard, as well as the expected performance of the alternatives, is unknown. For example, the standard might be an existing system that is being considered for replacement, but it is nevertheless simulated to provide a fair comparison with the alternatives. A second example occurs when the standard is the (known) least-cost system design, the design that will be implemented unless more expensive alternatives can significantly better its performance in terms of some measure other than cost. In the statistics literature this type of problem is known as "comparison with a control."

In other applications the performance of the standard may be considered known or certain (so that its variance is zero). An example is an existing system that has been in place long enough that its long-run average performance is well documented; the simulation might be undertaken to evaluate various upgrade strategies. A second example of a known standard is a goal or requirement—such as responding to customer calls within 30 minutes—when the purpose of the simulation study is to determine which system designs can meet or beat this standard. Clearly, a known standard is a limiting case of an unknown standard as the performance of the unknown standard becomes more and more certain. We unify the treatment of the known and unknown standard cases in this paper.

The procedures that we develop here share the following characteristics:

1. They require that the simulation experiment be performed in two stages, a first stage to assess the variability of the simulation output, and a second stage designed to achieve the desired precision of the comparison.

2. They exploit the concept of an *indifference zone*, which is an experimenter-specified difference in expected performance that is deemed practically significant, and therefore worth detecting.

3. They yield a decision, either that no alternative is better than the standard or that one or more of them is better. When at least one of the alternatives is better, the procedures indicate which alternative is the best. These decisions are guaranteed to be correct with an experimenter-specified probability.

4. They provide the following bounds: bounds on the difference between each alternative and the standard, when none of the alternatives betters it; and bounds on the difference between each alternative and the best of the others when one or more of the alternatives is better than the standard. These bounds are also guaranteed to be correct with at least an experimenter-specified probability.

Our procedures extend the work of Bechhofer and Turnbull (1978) and Paulson (1952) by adapting them to handle unequal variances and common random numbers—conditions frequently encountered in simulation experiments—and by adding the bounds on the differences. Like this earlier work, the specific procedures we derive depend on the simulation output data being normally distributed.

A characteristic of our procedures (and those of Bechhofer and Turnbull 1978 and Paulson 1952) is that they require one or more critical constants, constants that are neither easily calculated nor readily tabled. To remedy this problem, we also exhibit methods to *estimate* upper bounds on these constants, bounds that hold with an experimenter-specified probability. Because the probability that our procedures achieve their objectives is an increasing function of the critical constants, employing upper bounds ensures that the objectives are achieved with at least the prespecified probability.

The paper is organized as follows: The next section presents the generic comparison-with-a-standard procedure and establishes the key probability statements that guarantee its success. These key statements do not depend upon the output data being normally distributed. Section 3 customizes the generic procedure for different properties of the simulation output data; all of these customizations depend on normality. Section 4 shows how to estimate the critical values required by some versions of the comparison procedure. Section 5 presents a portion of an empirical evaluation of these procedures, while §6 explicitly illustrates one of them via a numerical example. Finally, §7 describes some direct extensions of the research. The longer proofs are contained in the online companion to this paper, available on the *Management Science* website at ⟨mansci.pubs.informs.org⟩.

2. Framework

In this section we precisely define the comparison problem of interest and put in place the framework that we use to derive procedures.

Let π_i denote the ith system, for $i = 0, 1, \ldots, k$, where π_0 is the designated benchmark or standard. Let X_{ij} represent the jth output (typically a sample average from within a replication or batch) from system i. In this paper $X_{i1}, X_{i2}, \ldots$ are taken to be independent and identically distributed (i.i.d.), a condition that is always true for replications, and is approximately true when batching within a single long replication if the underlying stochastic process is stationary and the batches are sufficiently large.

System i has expected performance $\mu_i = E[X_{ij}]$, and we denote the ordered but unknown means for alternative (nonstandard) systems $1, 2, \ldots, k$ as

$$\mu_{[1]} \le \mu_{[2]} \le \cdots \le \mu_{[k]}.$$

The system associated with $\mu_{[i]}$ is unknown, but is denoted $\pi_{[i]}$. The expected performance of the standard is denoted as μ_0. Without loss of generality, we assume that a larger expected value corresponds to better performance.

All of our procedures are based on estimating the true mean, μ_i, by a sample mean, so we let $\overline{X}_i$ denote the sample mean of all of the observations from system i, and let $\overline{X}_{[i]}$ denote the sample mean associated with the (nonstandard) system having mean $\mu_{[i]}$. The $\overline{X}_{[i]}$ need not be ordered, in contrast to $\overline{X}_{(i)}$, which denotes the ith smallest (nonstandard) sample mean; that is,

$$\overline{X}_{(1)} \le \overline{X}_{(2)} \le \cdots \le \overline{X}_{(k)}.$$

We wish to retain the standard when none of the alternatives is better, and we wish to select the best of the alternatives when it is at least a practically significant amount δ better than everything else. The parameter δ is called the *indifference-zone parameter*. If the best alternative is less than δ better than the other alternatives or the standard, then we are indifferent to which of these "good" systems we select. We will show that our procedures guarantee, with probability at least $1-\alpha$, that the selected system is within δ of the best *regardless of the configuration of the true means*. Of course, if the experimenter sets δ to be very small, the procedures may sometimes need to make fine distinctions between close competitors, and then a great deal of sampling may be required. In any case, as we will show, our goals will be achieved if the following hold:

$$\Pr\{\text{select } \pi_0\} \ge 1-\alpha \text{ whenever } \mu_0 \ge \mu_{[k]}. \quad (1)$$

$$\Pr\{\text{select } \pi_{[k]}\} \ge 1-\alpha \text{ whenever } \mu_{[k]} \ge \mu_{[k-1]} + \delta \text{ and } \mu_{[k]} \ge \mu_0 + \delta. \quad (2)$$

Our procedures provide a constant, c, and an algorithm to determine the number of outputs, N_i, to be obtained from π_i, such that (1) and (2) hold when we apply the following rule: Choose the standard if $\overline{X}_{(k)} \le \overline{X}_0 + c$; otherwise choose the alternative associated with $\overline{X}_{(k)}$. When the expected performance of the standard is known, then replace $\overline{X}_0$ by μ_0 in the rule.

A generic version of our procedure, where we implicitly assume that each system has finite variance, is as follows:

Generic Comparison-with-a-Standard Procedure

Step 0. Given k alternative systems and a standard, specify an initial sample size, n_0, an indifference-zone parameter δ, and a confidence level $1-\alpha$. Determine appropriate constants g and h, and let $c = \delta h/g$.

Step 1. Generate a sample $X_{i1}, X_{i2}, \ldots, X_{in_0}$ from system i, for $i = 0, 1, 2, \ldots, k$.

Step 2. Compute the appropriate variance estimator S_i^2 for each system i.

Step 3. Determine the required total sample size from system i as

$$N_i = \max\left\{n_0, \left\lceil \left(\frac{gS_i}{\delta}\right)^2 \right\rceil\right\} = \max\left\{n_0, \left\lceil \left(\frac{hS_i}{c}\right)^2 \right\rceil\right\}, \quad (3)$$

where $\lceil x \rceil$ denotes the least integer that is greater than or equal to x.

Step 4. Obtain additional outputs $X_{i,n_0+1}, X_{i,n_0+2}, \ldots, X_{i,N_i}$ from system i if needed, and compute the overall sample mean

$$\overline{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$$

for $i = 0, 1, 2, \ldots, k$.

Step 5. With confidence level greater than or equal to $1-\alpha$, apply the following rule:

Step 5a. If $\overline{X}_{(k)} \le \overline{X}_0 + c$, then choose the standard and form the one-sided joint confidence intervals

$$\mu_0 - \mu_i \le \overline{X}_0 - \overline{X}_i + c \tag{4}$$

for $i = 1, 2, \ldots, k$.

Step 5b. Otherwise, choose the alternative associated with the largest sample mean $\overline{X}_{(k)}$, and form the *multiple comparisons with the best (MCB)* confidence intervals

$$\mu_i - \max_{\ell \ne i} \mu_\ell \in \left[-\left(\overline{X}_i - \max_{\ell \ne i} \overline{X}_\ell - \delta\right)^-, \left(\overline{X}_i - \max_{\ell \ne i} \overline{X}_\ell + \delta\right)^+ \right] \tag{5}$$

for $i = 0, 1, 2, \ldots, k$, where $-x^- = \min\{0, x\}$, and $x^+ = \max\{0, x\}$.

Comment. If μ_0 is known, then we do not need to sample from system 0, and we replace $\overline{X}_0$ by μ_0 in Step 5.

Comment. If smaller expected performance is better, then simply multiply all of the observations by -1 before applying the procedure.

We now derive the fundamental probability statements that guarantee a correct selection. In this section our only requirement on X_{ij} is that the distribution of $X_{ij} - \mu_i$ is independent of μ_i; a sufficient condition is that the X_{ij} are normally distributed. We will use the notation "$\forall i_{a \ldots b}$" to mean $i = a, a+1, \ldots, b$.

The generic algorithm will ensure that (1) holds if

$$\Pr\left\{\overline{X}_i \le \overline{X}_0 + c, \forall i_{1 \ldots k}\right\} \ge 1 - \alpha \tag{6}$$

whenever $\mu_0 \ge \mu_i, \forall i_{1 \ldots k}$. Notice that (6) depends on the unknown means μ_i. However, if $\mu_0 \ge \mu_i, \forall i_{1 \ldots k}$, then

$$\begin{aligned} &\Pr\left\{\overline{X}_i \le \overline{X}_0 + c, \forall i_{1 \ldots k}\right\} \\ &\quad = \Pr\left\{(\overline{X}_i - \overline{X}_0) - (\mu_i - \mu_0) \le c - (\mu_i - \mu_0), \forall i_{1 \ldots k}\right\} \\ &\quad \ge \Pr\left\{(\overline{X}_i - \overline{X}_0) - (\mu_i - \mu_0) \le c, \forall i_{1 \ldots k}\right\}, \end{aligned} \tag{7}$$

and (7) does not depend on the unknown means. Similarly, (2) will hold if

$$\Pr\left\{\overline{X}_{[k]} > \overline{X}_0 + c, \overline{X}_{[k]} > \overline{X}_{[i]}, \forall i_{1 \ldots k-1}\right\} \ge 1 - \alpha \tag{8}$$

whenever $\mu_{[k]} \ge \mu_{[k-1]} + \delta$ and $\mu_{[k]} \ge \mu_0 + \delta$. Again, (8) depends on the unknown means μ_i. However, if $\mu_{[k]} \ge \mu_{[k-1]} + \delta$ and $\mu_{[k]} \ge \mu_0 + \delta$, then

$$\begin{aligned} &\Pr\left\{\overline{X}_{[k]} > \overline{X}_0 + c, \overline{X}_{[k]} > \overline{X}_{[i]}, \forall i_{1 \ldots k-1}\right\} \\ &\quad = \Pr\left\{(\overline{X}_{[k]} - \overline{X}_0) - (\mu_{[k]} - \mu_0) > c - (\mu_{[k]} - \mu_0), \right. \\ &\qquad \left. (\overline{X}_{[k]} - \overline{X}_{[i]}) - (\mu_{[k]} - \mu_{[i]}) > -(\mu_{[k]} - \mu_{[i]}), \forall i_{1 \ldots k-1}\right\} \\ &\quad \ge \Pr\left\{(\overline{X}_{[k]} - \overline{X}_0) - (\mu_{[k]} - \mu_0) > c - \delta, \right. \\ &\qquad \left. (\overline{X}_{[k]} - \overline{X}_{[i]}) - (\mu_{[k]} - \mu_{[i]}) > -\delta, \forall i_{1 \ldots k-1}\right\}, \end{aligned} \tag{9}$$

and (9) is independent of the unknown means. Therefore, we can attain (1) and (2) if we can derive a procedure that simultaneously guarantees that

$$\Pr\left\{(\overline{X}_i - \overline{X}_0) - (\mu_i - \mu_0) \le c, \forall i_{1 \ldots k}\right\} \ge 1 - \alpha \tag{10}$$

$$\Pr\left\{(\overline{X}_{[k]} - \overline{X}_0) - (\mu_{[k]} - \mu_0) > c - \delta, (\overline{X}_{[k]} - \overline{X}_{[i]}) - (\mu_{[k]} - \mu_{[i]}) > -\delta, \forall i_{1 \ldots k-1}\right\} \ge 1 - \alpha. \tag{11}$$

When μ_0 is known, then we replace $\overline{X}_0$ by μ_0 in (10) and (11).

We derive procedures for a variety of cases in §3, where we assume that the simulation output data are normally distributed. However, (10) and (11) depend only on the weaker condition that the distribution of $X_{ij} - \mu_i$ is independent of μ_i. Therefore, procedures for other types of data could be based on satisfying (10) and (11) provided appropriate constants g and h can be determined.

We have shown that (10) and (11) guarantee a correct selection with probability $\ge 1 - \alpha$. The following theorem shows that (10) and (11) are also sufficient to establish the confidence intervals in Step 5 of the generic procedure:

Theorem 1. *If* (10) *and* (11) *hold and the distribution of* $X_{ij} - \mu_i$ *is independent of* μ_i, *then the events* (4) *and* (5) *occur individually with probability greater than or equal to* $1 - \alpha$.

Proof. That (4) holds with probability greater than or equal to $1 - \alpha$ follows immediately from (10). To show that (5) holds, we first assume that $\mu_{[k]} \ge \mu_{[k-1]} + \delta$ and $\mu_{[k]} \ge \mu_0 + \delta$. Then, since $c > 0$, we have

$$\begin{aligned} &\Pr\left\{\overline{X}_{[k]} > \overline{X}_0, \overline{X}_{[k]} > \overline{X}_{[i]}, \forall i_{1 \ldots k-1}\right\} \\ &\quad \ge \Pr\left\{\overline{X}_{[k]} > \overline{X}_0 + c, \overline{X}_{[k]} > \overline{X}_{[i]}, \forall i_{1 \ldots k-1}\right\} \\ &\quad \ge 1 - \alpha, \end{aligned}$$

where the last inequality follows because (11) implies (8). Therefore, (5) holds with probability greater than

or equal to $1-\alpha$ by Theorem 1 of Nelson and Matejcik (1995). □

As a consequence of our Theorem 1, we are guaranteed that the mean of the system with the largest overall sample mean (be it the standard, or one of the alternatives) is within δ of the largest true mean with probability greater than or equal to $1-\alpha$ *under all possible configurations of* $\mu_0, \mu_1, \ldots, \mu_k$. To state the following corollary, let B denote the index of the system with the largest overall sample mean.

COROLLARY 1. *If* (10) *and* (11) *hold and the distribution of* $X_{ij}-\mu_i$ *is independent of* μ_i, *then*

$$\Pr\left\{\mu_B - \max_{\ell\neq B}\mu_\ell \geq -\delta\right\} \geq 1-\alpha.$$

PROOF. From Theorem 1, we know that

$$\mu_B - \max_{\ell\neq B}\mu_\ell \in \left[-\left(\overline{X}_B - \max_{\ell\neq B}\overline{X}_\ell - \delta\right)^-, \left(\overline{X}_B - \max_{\ell\neq B}\overline{X}_\ell + \delta\right)^+\right]$$

occurs with probability greater than or equal to $1-\alpha$. But since $\overline{X}_B \geq \overline{X}_\ell, \forall \ell_{0\ldots k}$, the smallest possible value of the lower bound is $-\delta$. □

Notice that we may still select the standard even if $\overline{X}_0$ is not the largest sample mean, because our requirements (1) and (2) favor the standard, seeking to retain it even if it is tied with the best. Thus, we require substantial evidence before giving up the standard. Corollary 1 guarantees that we get a "good" system, with high probability, if we select the one with the largest sample mean.

3. Procedures

We have derived specific instances of the generic comparison-with-a-standard procedure to handle the types of data encountered in simulation. Specifically, we consider the following cases:

Status of the Standard. We consider the case in which μ_0 is known and when it is unknown and must be estimated along with the expected performance of the alternatives. When μ_0 is unknown, it may be estimated via a simulation experiment or by collecting data on the real system.

Unequal Variances Across Systems. All of our procedures permit unequal variances across systems, although Case C is an approximation when the variances are not all the same. This is one of the contributions of our work beyond that of Bechhofer and Turnbull (1978) and Paulson (1952).

Dependence Across Systems. We develop procedures in which all systems are simulated (or sampled, if π_0 is a real system) independently. For the case of unknown μ_0, we also develop procedures in which the simulations of all systems are driven by common random numbers (CRN), inducing dependence across systems. One way we account for the effect of CRN is to assume that the induced variance-covariance matrix across systems satisfies a condition known as *sphericity*. In brief, assuming sphericity leads to a procedure that approximates the variance of the difference between observations from any two of the systems by the average variance of the difference between observations from all pairs of systems. See Nelson (1993) and Nelson and Matejcik (1995) for further discussion of the implications of sphericity, as well as empirical tests for it.

We also show that it may be counterproductive to use CRN when μ_0 is known, or μ_0 is unknown and estimated independently of the alternatives. Therefore, we do not derive CRN-compatible procedures for these cases.

For readers only interested in applications, §3.1 gives the essential information required to customize the generic procedure for various cases. The proofs are in the online companion.

3.1. Customizing the Generic Procedure

All of our procedures require critical values (g, h) that satisfy

$$\Pr\{H \leq h\} = 1-\alpha,$$
$$\Pr\{G \leq g\} = 1-\alpha,$$

where H and G are random variables whose distributions depend on whether or not we know μ_0, we use CRN, or we assume sphericity. In addition, all of the procedures require a variance estimator S_i^2 associated with system i from the first-stage sample (Step 2). In this section we collect the definitions of G, H, and S_i^2 for each of the cases considered below. To do so,

let $Z_0, Z_1, \ldots, Z_k$ be i.i.d. $N(0,1)$ random variables; let $Y_0, Y_1, \ldots, Y_k$ be i.i.d. χ^2 random variables, each with $n_0 - 1$ degrees of freedom and independent of the Z_is; let $T, T_1, T_2, \ldots, T_k$ be i.i.d. t random variables, each with $n_0 - 1$ degrees of freedom; and let $(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k)$ be a multivariate-t random vector with common correlation $1/2$ and $k(n_0 - 1)$ degrees of freedom. In the formulas below a $\cdot$ subscript indicates averaging with respect to that subscript for the n_0 observations from the first stage of sampling.

Case A. μ_0 known, alternative systems simulated independently, the X_{ij} are normally distributed, and the variances across systems may be unequal.

$$\begin{aligned} H &= \max\{T_1, T_2, \ldots, T_k\}; \\ G &= \max\{Z_k[(n_0-1)/Y_k]^{1/2} \\ &\quad + h, Z_i[(n_0-1)(1/Y_i + 1/Y_k)]^{1/2}, \forall i_{1\cdots k-1}\}; \\ S_i^2 &= \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}(X_{ij} - \overline{X}_{i\cdot})^2. \end{aligned}$$

Case B. μ_0 unknown, all systems simulated independently, the X_{ij} are normally distributed, and the variances across systems may be unequal.

$$\begin{aligned} H &= \max\{Z_i[(n_0-1)(1/Y_i + 1/Y_0)]^{1/2}, \forall i_{1\cdots k}\}; \\ G &= \max\{Z_0[(n_0-1)(1/Y_0 + 1/Y_k)]^{1/2} \\ &\quad + h, Z_i[(n_0-1)(1/Y_i + 1/Y_k)]^{1/2}, \forall i_{1\cdots k-1}\}; \\ S_i^2 &= \frac{1}{n_0 - 1}\sum_{j=1}^{n_0}(X_{ij} - \overline{X}_{i\cdot})^2. \end{aligned}$$

Case C. μ_0 unknown, all systems simulated using CRN, and $(X_{0j}, X_{1j}, \ldots, X_{kj})$ have a multivariate normal distribution whose variance-covariance matrix satisfies sphericity.

$$\begin{aligned} H &= \max\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}; \\ G &= \max\{\mathcal{T}_1 + h, \mathcal{T}_2, \ldots, \mathcal{T}_k\}; \\ S_i^2 &= S^2 = \frac{2}{k(n_0-1)}\sum_{\ell=0}^{k}\sum_{j=1}^{n_0}(X_{\ell j} - \overline{X}_{\ell\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot})^2. \end{aligned}$$

Case D. μ_0 unknown, all systems simulated using CRN, and $(X_{0j}, X_{1j}, \ldots, X_{kj})$ have a multivariate normal distribution whose variance-covariance matrix is unknown. For Case D, whose proof exploits the Bonferroni inequality, it is easier to provide somewhat different definitions of g and h; in particular, (g, h) simultaneously solve

$$\begin{aligned} &1 - k\Pr\{T > h\} = 1 - \alpha; \\ &1 - \Pr\{T + h > g\} - (k-1)\Pr\{T > g\} = 1 - \alpha \end{aligned}$$

(such a solution exists and is unique). The variance estimator for Case D is

$$S_i^2 = \widehat{S}^2 = \max_{i,\,\ell:i\neq\ell}\left\{\frac{1}{n_0 - 1}\sum_{j=1}^{n_0}\left[(X_{ij} - X_{\ell j}) - (\overline{X}_{i\cdot} - \overline{X}_{\ell\cdot})\right]^2\right\},$$

the largest variance of the difference.

Notice that Cases C and D use a common sample size N for all systems. In §4 we provide a procedure for estimating the quantiles (g, h) for Cases A, B, and C. For Case D, a simple numerical search suffices.

3.2. The Issue of Normality

Cases A–D all assume that the simulation output data are normally distributed, either marginally or jointly. Normality of the first-stage observations is important because the joint distribution of $\overline{X}_i$ and S_i^2 is central to the derivations of the procedures. Therefore, it makes sense to consider whether or when this is a reasonable expectation for simulation output data.

In many simulation studies the basic output data $X_{i1}, X_{i2}, \ldots$ are themselves averages of large numbers of other more basic outputs. For example, if $X_{i1}, X_{i2,}, \ldots$ correspond to different replications, and the performance measure is expected cycle time for a product, then X_{ij} would typically be the average of the cycle times for a large number (perhaps hundreds) of individual products that were completed during the jth replication. In this case, the central limit theorem suggests that approximate normality of the X_{ij} may be anticipated.

Clearly situations do arise in which the output from each replication is not even approximately normally distributed. For instance, if each replication produces a single observation of time to failure for a system, then there is no a priori reason to expect normality. However, if a large number of replications can be obtained, say m, then Goldsman et al. (1991) suggest that they be partitioned into n_0 "macroreplications," each consisting of m/n_0 "microreplications." The average value within each macroreplication is then treated as the basic output data value. If m is large enough, then the microreplication averages will be approximately normally distributed.

Even when only a single replication is obtained in order to estimate long-run performance in a steady-state simulation, we may anticipate that the

$X_{i1}, X_{i2}, \ldots$ will be approximately normally distributed if they correspond to nonoverlapping batch means of many individual observations. Batch-size analysis in Matejcik and Nelson (1995) then suggests that the number of batches be kept to roughly 40, so that each batch mean is based on a very large number of observations. This same guideline applies to the number of macroreplications that should be formed when outputs are obtained across replications.

In §5, we present empirical examples to investigate the robustness of our procedures to departures from normality.

3.3. CRN May Be Counterproductive

We have not presented procedures that incorporate CRN for the simulation of the alternatives $\pi_1, \pi_2, \ldots, \pi_k$ when π_0 is simulated or sampled independently, or when μ_0 is known so that π_0 is not simulated at all. In this section we present a brief analysis that shows why CRN may be counterproductive in these cases, and it is therefore safer to simulate the alternatives independently.

Suppose that we have $k = 2$ alternatives, μ_0 is known, and we take exactly one observation from each of the alternative systems, yielding data (X_1, X_2). Suppose further that (X_1, X_2) has a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. We assume $\rho > 0$, representing the effect of CRN.

For this simple example, the lower bound on PCS (11) becomes

$$\begin{aligned}\Pr\{X_{[2]} - \mu_{[2]} > c - \delta, (X_{[2]} - X_{[1]}) - (\mu_{[2]} - \mu_{[1]}) > -\delta\} \\ = \Pr\{X_{[2]} - \mu_{[2]} < \delta - c, (X_{[2]} - X_{[1]}) \\ - (\mu_{[2]} - \mu_{[1]}) < \delta\}. \qquad (12)\end{aligned}$$

Basic mathematical statistics shows that

$$\gamma = \mathrm{Cov}[(X_{[2]} - \mu_{[2]}), (X_{[2]} - X_{[1]}) - (\mu_{[2]} - \mu_{[1]})] < 0$$

if $\rho > \sigma_{[2]}/\sigma_{[1]}$ (where $\sigma_{[i]}^2$ is the variance of system $\pi_{[i]}$); this is certainly possible when the variances are unequal. However, by Slepian's inequality (see, for instance, Tong 1980), we know that (12) is an increasing function of γ. Therefore, when $\rho > \sigma_{[2]}/\sigma_{[1]}$, the lower bound on the probability of correct selection is *larger* if the alternatives are simulated independently (in which case $\rho = 0$ and $\gamma > 0$) rather than with CRN.

Intuitively, when the standard is fixed and all of the alternatives hang together (due to CRN), then if one of the alternatives is difficult to distinguish from the standard, they all are. A similar argument holds when μ_0 is unknown, π_0 is simulated or sampled independently of the alternatives, and the alternatives are simulated using CRN.

4. Estimating Critical Values

Traditionally, critical values for statistical inference have been computed, often via intensive numerical integration, and then tabled for later use. This becomes impractical when the desired critical values depend on a large number of problem parameters. In the present setting, the critical values (g, h) depend on the confidence level, $1 - \alpha$, the number of systems, k, the first-stage sample size, n_0, and whether or not CRN is employed. When the required numerical integration is of low dimension, then real-time numerical calculation of the critical values may be possible. However, each problem type may then require a finely tuned numerical procedure that works well over the feasible range of problem parameters.

As computation speed increases, another approach becomes viable: Use a separate simulation experiment to *estimate* the critical values as needed for the problem at hand. As we will show, the present context is ideal for this approach.

Cases A, B, and C presented in §3 require a pair of quantiles (g, h) that satisfy

$$\Pr\{(\overline{X}_i - \overline{X}_0) - (\mu_i - \mu_0) \le c, \forall i_{1\cdots k}\} \ge \Pr\{H \le h\} = 1 - \alpha,$$

which corresponds to the Probability Requirement (10), and

$$\begin{aligned}\Pr\{(\overline{X}_{[k]} - \overline{X}_0) - (\mu_{[k]} - \mu_0) \\ > c - \delta, (\overline{X}_{[k]} - \overline{X}_{[i]}) - (\mu_{[k]} - \mu_{[i]}) \\ > -\delta, \forall i_{1\cdots k-1}\} \ge \Pr\{G \le g\} = 1 - \alpha,\end{aligned}$$

which corresponds to the Probability Requirement (11). Notice that for Cases A–C, H and G are continuous random variables; and we can write $G = \max\{M_1 + h, M_2, \ldots, M_k\}$, where $M_1, M_2, \ldots, M_k$ are random variables that are easily sampled. (See the online companion for additional details, in particular

on how G and H correspond to the probability statements above.) Further, for all of our procedures the total sample size from system i satisfies

$$N_i \geq \left(\frac{gS_i}{\delta}\right)^2$$

so that N_i is an increasing function of g. This implies that we would prefer an estimate of g, say $\hat{g}$, that is a bit larger than g, rather than one that is a bit smaller than g—because if N_i is too small, then we may not achieve the desired probability of correct selection and confidence interval coverage, while N_i too large makes the procedure conservative. We cannot guarantee that our simulation estimate $\hat{g}$ is greater than or equal to g, but we can find an estimator $\hat{g}$ with the property that

$$\Pr\{\hat{g} < g\} \leq \beta$$

for some small, prespecified value of β. In this section we show that the following procedure yields such an estimator:

Procedure (g, h) **Bound**

Step 1. Select positive integers m_h and m_g, and confidence levels β_h and β_g such that $\beta_h + \beta_g = \beta$.

Step 2. Find the smallest integers u_h and u_g such that

$$\sum_{\ell=u_h}^{m_h} \binom{m_h}{\ell} (1-\alpha)^\ell \alpha^{m_h-\ell} \leq \beta_h;$$
$$\sum_{\ell=u_g}^{m_g} \binom{m_g}{\ell} (1-\alpha)^\ell \alpha^{m_g-\ell} \leq \beta_g.$$

Step 3. Generate $H_1, H_2, \ldots, H_{m_h}$, i.i.d. copies of H, and set $\hat{h} = H_{(u_h)}$.

Step 4. Generate $\widehat{G}_1, \widehat{G}_2, \ldots, \widehat{G}_{m_g}$, i.i.d. copies of $\widehat{G}$, where $\widehat{G} = \max\{M_1 + \hat{h}, M_2, \ldots, M_k\}$.

Step 5. Return $\hat{h}$ and $\hat{g} = \widehat{G}_{(u_g)}$.

To prove that the procedure works, we address a somewhat more general case: For continuous cdfs F_H and F_G, define $h = F_H^{-1}(1-\alpha_h)$ and $g = F_G^{-1}(1-\alpha_g; h)$. Suppose that F_G^{-1} is a nondecreasing function of the parameter h. Suppose also that we have a procedure $\mathcal{Q}_h$ that provides an estimator $\hat{h}$ with the property that $\Pr\{\hat{h} < h\} \leq \beta_h$. Further, we have a procedure $\mathcal{Q}_g(\hat{h})$ that takes as input $\hat{h}$, and returns as output an estimator $\hat{g}$ with the property that

$$\Pr\{\hat{g} < g | \hat{h} \geq h\} \leq \beta_g.$$

THEOREM 2. *If $\hat{g}$ is defined by procedures $\mathcal{Q}_h$ and $\mathcal{Q}_g(\hat{h})$, then* $\Pr\{\hat{g} < g\} \leq \beta$.

PROOF.

$$\begin{aligned}\Pr\{\hat{g} < g\} &= \Pr\{\hat{g} < g | \hat{h} \geq h\}\Pr\{\hat{h} \geq h\} \\ &\quad + \Pr\{\hat{g} < g | \hat{h} < h\}\Pr\{\hat{h} < h\} \\ &\leq \beta_g + \beta_h = \beta, \qquad (13)\end{aligned}$$

where (13) follows from properties of $\mathcal{Q}_h$ and $\mathcal{Q}_g(\hat{h})$. □

It is straightforward to verify that Procedure (g, h) Bound satisfies the conditions of Theorem 2, since

$$\begin{aligned}\Pr\{H_{(u_h)} < h\} &= \Pr\{\#\{H_i \leq h\} \geq u_h\} \\ &= \sum_{\ell=u_h}^{m_h} \binom{m_h}{\ell} (1-\alpha)^\ell \alpha^{m_h-\ell} \\ &\leq \beta_h,\end{aligned}$$

where the second line follows from the definition of h and the last line follows from our choice of u_h. A similar argument holds for $\Pr\{\widehat{G}_{(u_g)} < g | \hat{h} \geq h\}$.

COMMENT. Many ranking, selection, and multiple comparison procedures, and in particular the procedures of Nelson and Matejcik (1995) and Matejcik and Nelson (1995) for selecting the best, require a single critical value similar to our h. Thus, a one-sided upper confidence interval for h that holds with probability greater than or equal to $1-\beta_h$ can be achieved by stopping at Step 3 in Procedure (g, h) Bound.

EXAMPLE. Suppose that we want an overall confidence level of $1-\alpha = 0.9$ and a 95% upper confidence bound on the critical value, implying that $\Pr\{\hat{g} < g\} \leq \beta = 0.05$. If we set $\beta_h = \beta_g = 0.025$ and $m_h = m_g = 1{,}000$ observations, then to three decimal places

$$\sum_{\ell=919}^{1000} \binom{1000}{\ell} (0.9)^\ell (0.1)^{1000-\ell} = 0.023 \leq 0.025$$

so that $u_h = u_g = 919$.

Tables 1 and 2 give (g, h) values when the number of first-stage observations is 10 and the desired overall confidence level is 0.90 or 0.95, respectively, for various numbers of alternative systems k. Notice that even with as few as 20,000 observations, the 99% upper confidence bounds are quite close to

Table 1 **Critical Values h above g for $n_0 = 10$ and $1 - \alpha = 0.90$**

	Case			
k	A	B	C	D
2	1.817	2.588 (2.637)	1.650 (1.680)	1.833
	3.345 (3.381)	4.652 (4.697)	3.026 (3.063)	3.251
3	2.064	2.922 (2.966)	1.786 (1.814)	2.086
	3.622 (3.660)	5.023 (5.070)	3.156 (3.195)	3.514
4	2.238	3.158 (3.200)	1.881 (1.905)	2.262
	3.860 (3.894)	5.240 (5.291)	3.253 (3.285)	3.697
5	2.373	3.299 (3.354)	1.959 (1.991)	2.398
	4.010 (4.047)	5.375 (5.427)	3.304 (3.338)	3.837

Note. When the critical value is estimated via simulation, a 99% upper confidence bound is given next to the estimate in parentheses. Estimates are based on 20,000 observations.

Table 2 **Critical Values h above g for $n_0 = 10$ and $1 - \alpha = 0.95$**

	Case			
k	A	B	C	D
2	2.254	3.182 (3.251)	2.055 (2.102)	2.262
	4.174 (4.222)	5.850 (5.918)	3.855 (3.900)	4.112
3	2.499	3.506 (3.585)	2.147 (2.188)	2.510
	4.437 (4.489)	6.216 (6.289)	3.924 (3.965)	4.366
4	2.673	3.740 (3.800)	2.228 (2.264)	2.685
	4.648 (4.702)	6.422 (6.501)	3.968 (4.007)	4.545
5	2.809	3.852 (3.924)	2.234 (2.358)	2.821
	4.829 (4.868)	6.569 (6.643)	4.047 (4.087)	4.684

Note. When the critical value is estimated via simulation, a 99% upper confidence bound is given next to the estimate in parentheses. Estimates are based on 20,000 observations.

the point estimates. One way to choose m_g and m_h is to increase them until the estimates $\hat{g}$ and $\hat{h}$ are sufficiently close to their upper confidence bounds. S-Plus code (MathSoft, Inc.) to obtain critical values for all four cases can be obtained from ⟨http://www.iems.nwu.edu/~nelsonb/NSF/⟩.

5. Empirical Evaluation

To evaluate the robustness of the Case C version of the procedure to departures from sphericity, and to evaluate the conservatism of the Case D version of the procedure in general, we performed an empirical study. Since it is not possible to control the extent to which system-simulation examples depart from sphericity, we focused instead on the space of normally distributed output vectors with nonnegative covariances (the assumed effect of CRN). We estimated the probability of correct selection (PCS) over this space, but did not estimate confidence-interval coverage separately since it is implied by the correct-selection guarantee.

We considered only the equal means configuration (EMC), $\mu_0 = \mu_1 = \cdots = \mu_k$, and the slippage configuration (SC), $\mu_k - \delta = \mu_{k-1} = \mu_{k-2} = \cdots = \mu_0$, because the minimum PCS should occur at these configurations. In the EMC, a "correct selection" means retaining the standard, while in the SC it means selecting system k.

We fixed $\delta = \frac{1}{2}$, 1, and 2 in units of the average standard error of the first-stage sample means; specifically, $\delta = (1/2\sqrt{n_0})$, $(1/\sqrt{n_0})$, or $(2/\sqrt{n_0})$, where the average variance of an observation across all $k+1$ systems was always fixed to be 1. When $\delta = \frac{1}{2}$ there will be a large second-stage sample; $\delta = 1$ implies that there will usually be a modest second-stage sample; while $\delta = 2$ implies that second-stage sampling is rarely required.

In addition to varying δ, we considered different configurations of the systems' variances, $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2$, where σ_i^2 is the variance of an observation from system i. Specifically, we considered equal variances across all systems, the best system having 20% larger variance than all other systems, and the best system having 20% smaller variance than all other systems (in the EMC the best system is the standard, while in the SC the best system is system k). We chose not to investigate drastically unequal variances because comparisons based on mean performance only make sense when differences in variances do not dominate differences in means. Note, however, that the Case D procedure is valid no matter what the variances are. In all cases $(k+1)^{-1} \sum_{i=0}^{k} \sigma_i^2 = 1$. Finally, to assess the impact of nonnormality, we also generated data from lognormal distributions whose skewness and kurtosis (standardized third and fourth moments) differed from those of the normal distribution.

The experiments were conducted as follows:

1. Fix the number of systems, k, number of first-stage replications from each system, n_0, and confidence level $1 - \alpha$. We considered $k = 3$ and 5 systems

(which implies $k+1=4$ or 6 systems, including the standard), $n_0=10$ replications, and $1-\alpha=0.95$. Fix $\mathbf{D}$, a $(k+1)\times(k+1)$ matrix with off-diagonal elements 0 and diagonal $(\sigma_0, \sigma_1, \ldots, \sigma_k)$.

2. Generate a random k-dimensional correlation matrix $\boldsymbol{\Xi}$ using the method of Marsaglia and Olkin (1984). This method transforms a randomly generated point on the k-dimensional unit sphere into a correlation matrix. We modified the method to generate a point on the unit sphere with all nonnegative coordinates, which leads to a correlation matrix with all nonnegative elements. Set $\boldsymbol{\Sigma}=\mathbf{D}\boldsymbol{\Xi}\mathbf{D}$ to obtain a covariance matrix with variances $(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$ and implied correlation matrix $\boldsymbol{\Xi}$.

3. Generate n_0 i.i.d. random vectors $\mathbf{X}_j \sim$ (distributed as) $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, for $j=1, 2, \ldots, n_0$.

4. Compute the total sample size $N_i=\max\{n_0, \lceil(gS_i/\delta)^2\rceil\}$, where g and S_i depend on the procedure being evaluated. (For Cases C and D, $N_0=N_1=\cdots=N_k=N$.)

5. Generate $N-n_0$ i.i.d. random vectors $\mathbf{X}_j \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, for $j=n_0+1, n_0+2, \ldots, N$.

6. (a) In the EMC, our simulation sets the mean for every system to zero. Here we wish to select the standard, and so we score a correct selection if $\{\overline{X}_0+c>\overline{X}_i, \forall i_{1\ldots k}\}$.

(b) In the SC, our simulation sets the mean for system π_k exactly δ higher than all of the others. Thus, we score a correct selection if $\{\overline{X}_k+\delta>\overline{X}_i, \forall i_{1\ldots k-1};\ \overline{X}_k+\delta>\overline{X}_0+c\}$.

7. Repeat Steps 3–6 a total of 2,000 times to obtain an estimate of PCS for the covariance matrix $\boldsymbol{\Sigma}$ (2,000 replications give two significant digits of precision).

8. Repeat Steps 2–7 a total of 5,000 times to estimate the distribution of PCS over the space of covariance matrices, $\boldsymbol{\Sigma}$.

The experiments bypass one problem that affects all parametric multiple-comparison procedures—nonnormal data—and instead focus on the effect of positive correlation and unequal variances. The results are therefore optimistic in the same way that any parametric multiple-comparison procedure is optimistic with regard to the normality assumption. The results are pessimistic in the sense that we seldom encounter the EMC or SC in practice.

Before presenting some illustrative results, we summarize our conclusions from the complete set of experiments:

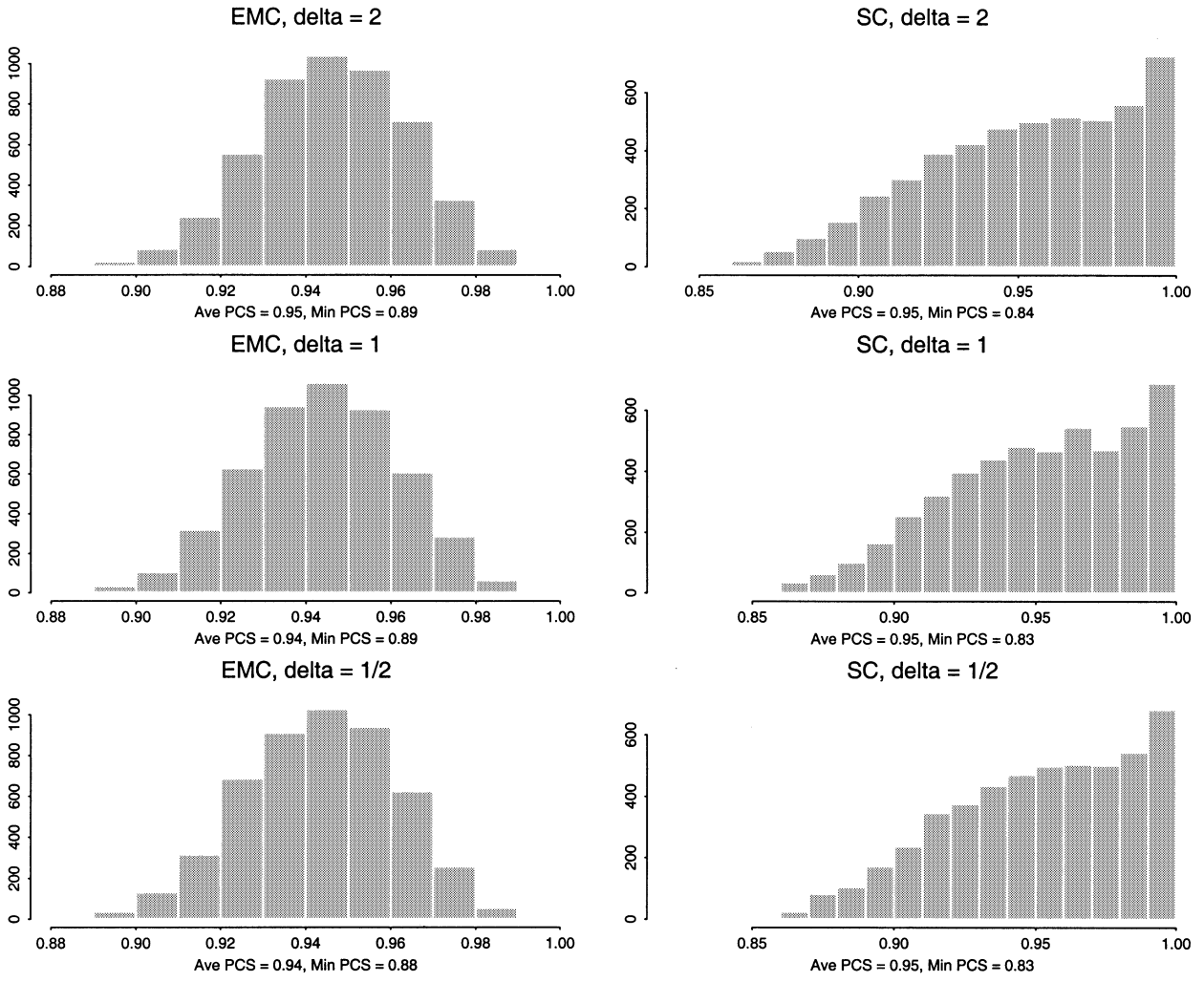
- Case C, the procedure based on assuming sphericity, achieved an *average* PCS of approximately 0.95 across the 5,000 randomly generated covariance matrices for each combination of k, δ, and $(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$ considered. However, probabilities of correct selection as low as 0.83, although rare, were observed in the SC because *the standard was selected when the best was in fact system k*. This is less robust performance than that observed by Nelson and Matejcik (1995) for a similar two-stage procedure designed only to select the best. *Therefore, the Case C procedure performs as desired on average, but has a higher than advertised risk of retaining the standard when one of the alternatives is exactly δ better than the standard.*
- Case D, the procedure based on the Bonferroni inequality, achieved a PCS of at least 0.95 for *each* randomly generated covariance matrix over all combinations of k, δ, and $(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$ considered. This was expected, since the procedure has been proven to achieve the PCS for normally distributed data. However, the average PCS can be significantly higher than 0.95, particularly as k is increased from 3 to 5 systems. *Therefore, the Case D procedure assures that the desired PCS is attained at the risk of delivering a higher than requested PCS by taking a larger total sample than is actually required.*
- Neither δ nor $(\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$ had a noticeable effect on the results. *Therefore, neither procedure is significantly affected by mild differences in systems' variances.*

Figures 1 and 2 present illustrative results for the Case C and D procedures, respectively. Each histogram summarizes the estimated PCS for 5,000 randomly generated covariance matrices, and there is one histogram for each value of δ and each configuration of the means. Common random numbers were employed in each figure.

In Figure 1, consider the SC with $\delta=2$. Although the average PCS is 0.95, the minimum observed is 0.84, and there is a 7% chance of PCS less than 90%. This illustrates the risk in using Case C, with the reward being greatly reduced sample sizes.

In Figure 2, consider the EMC with $\delta=\frac{1}{2}$. Notice that the average PCS is 0.99, and the minimum

Figure 1 **Estimated Probability of Correct Selection for Case C with $k = 5$ Systems and $n_0 = 10$ Replications over the Space of Randomly Generated Σ with Equal Variances**



observed PCS is 0.98, even though the nominal PCS is 0.95. This illustrates the conservatism inherent in using a procedure based on the Bonferroni inequality.

To assess the impact of nonnormal data on the procedures, we applied the Case C and D versions to lognormally distributed data with increasing levels of skewness and kurtosis, relative to the normal distribution (which has skewness 0 and kurtosis 3). The data were generated by transforming multivariate normal data obtained as described above; specifically, if X is normal, then e^X is lognormally distributed. Therefore, the data are still positively correlated across systems, but with correlations slightly altered by the exponential transformation. Parameters of the underlying normal distribution were chosen to obtain the desired skewness and kurtosis, then the distribution was shifted and scaled to place the means in the EMC or SC and make the variances 1. In every other respect the experiments were conducted as described above.

Table 3 shows the average of the 5,000 estimated PCS values for three lognormal models, with the corresponding normal model included for comparison, for Case C with $k = 5$. When skewness and kurtosis differ somewhat from normality (1.78, 9.10), the procedure still maintains an average PCS near 0.95.

Figure 2 **Estimated Probability of Correct Selection for Case D with $k = 3$ Systems and $n_0 = 10$ Replications over the Space of Randomly Generated Σ with Equal Variances**

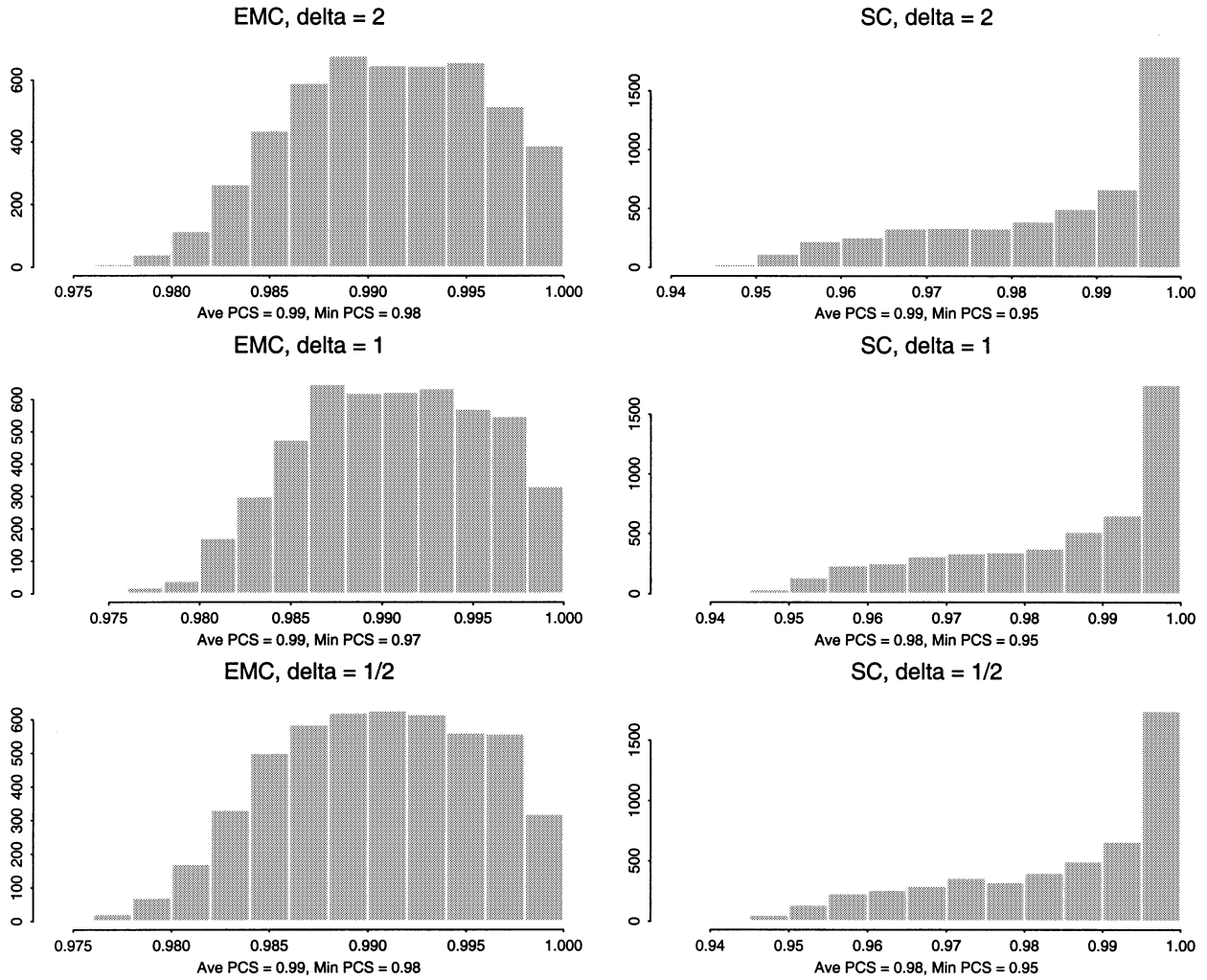


Table 3 **The Effect of Nonnormality on Average PCS for Case C**

		$\delta = 2$		$\delta = 1$		$\delta = 1/2$	
Distribution	(Skewness, Kurtosis)	EMC PCS	SC PCS	EMC PCS	SC PCS	EMC PCS	SC PCS
normal	(0, 3)	0.95	0.95	0.94	0.95	0.94	0.95
lognormal	(1.78, 9.10)	0.94	0.95	0.93	0.94	0.92	0.94
lognormal	(4.00, 41.00)	0.94	0.95	0.89	0.92	0.87	0.91
lognormal	(6.168, 113.173)	0.95	0.96	0.87	0.91	0.84	0.89

Note. In all cases $k = 5$, $\sigma_i^2 = 1$ for all i, $n_0 = 10$, δ is measured in units of $1/\sqrt{n_0}$, and nominal PCS is 0.95.

However, as the departure becomes more dramatic, the average PCS drops below the nominal level. The PCS values stayed closer to 0.95 when $k = 3$. Thus, the procedure should be applied with caution when data are expected or known to differ substantially from the normal model; mild departures, however, should present no difficulty.

Table 4 shows the average of the 5,000 estimated PCS values for the same three lognormal models for Case D with $k = 3$. In no case did the *average* PCS drop below 0.95. However, unlike the normal case, for certain correlation matrices the estimated PCS was as

Table 4 The Effect of Nonnormality on Average PCS for Case D

		$\delta = 2$		$\delta = 1$		$\delta = 1/2$	
		EMC	SC	EMC	SC	EMC	SC
Distribution	(Skewness, Kurtosis)	PCS	PCS	PCS	PCS	PCS	PCS
normal	(0, 3)	0.99	0.99	0.99	0.98	0.99	0.98
lognormal	(1.78, 9.10)	0.99	0.98	0.98	0.98	0.98	0.98
lognormal	(4.00, 41.00)	0.99	0.98	0.97	0.97	0.96	0.96
lognormal	(6.168, 113.173)	0.98	0.98	0.96	0.96	0.95	0.95

Note. In all cases $k = 3$, $\sigma_i^2 = 1$ for all i, $n_0 = 10$, δ is measured in units of $1/\sqrt{n_0}$, and nominal PCS is 0.95.

low as 0.87, although this was very rare. Performance was better (less chance of being below the nominal PCS) when $k = 5$.

6. Example

The following question arose in research on agile manufacturing systems. Suppose a portion of such a system consists of two stations in tandem, Station 1 and Station 2, but just *one* operator. Jobs come into the system at Station 1 at the rate of λ per hour. After arrival a job needs to be set up at the station by the operator. After setting up the job, the station processes it to completion *without requiring any assistance by the operator.* Following completion at Station 1, the job needs to be set up by the operator at Station 2, and then is processed at Station 2 without requiring any assistance from the operator. After completion at Station 2, the job leaves the system. The processing rate and the set-up rate at the two machines are μ_1, β_1 and μ_2, β_2, respectively. Assume that there is no walking time between the two stations. Holding costs of c_1 and c_2 per job per unit time are incurred while the jobs are in Stations 1 and 2, respectively. The operator has to decide which station to set up first when there are jobs waiting at both stations. The question is, what policy for the operator minimizes the expected holding cost of the system? (For further details, see Nelson and Banerjee 2001.)

We considered the following seven policies, some of which have been examined previously in Desruelle and Steudel (1996) and Nakade et al. (1997): The worker sets up the jobs on a first-come-first-serve basis (FIFO); this policy is considered to be the standard. The worker gives priority to Station 1 (SEIZE 1) or Station 2 (SEIZE 2). The worker immediately sets up any job that needs it at Station 1 (PREEMPT1) or at Station 2 (PREEMPT2), preempting any set up they are doing at the other station. Or, finally, the operator follows the FIFO policy until the number of jobs in Buffer 1 (respectively, 2) reaches n, at which point she switches to the PREEMPT1 (respectively, PREEMPT 2) policy, and switches back to FIFO when the number of jobs in the buffer falls below n. For this example we used $n = 3$. These policies are called TH1(3) and TH2(3), respectively.

Table 5 Service and Cost Parameters for the Agile Manufacturing Example

Station	Processing rate (jobs/hr)	Set-up rate (jobs/hr)	Holding cost
Station 1	4	4	\$1/job/hr
Station 2	6	6	\$1/job/hr

Interarrival, service, and set-up distributions are taken to be exponential. The arrival rate is fixed at $\lambda = 2$ jobs/hour; the service distribution parameters are shown in Table 5 with the holding costs. Notice that with the cost in both buffers being equal, minimization of the total holding cost is equivalent to the minimization of the total work in process in the system; our goal is to find the policy with the lowest expected holding cost per unit time, if it is lower than the standard (FIFO).

A cost reduction of more than \$1 was considered significant, so we set $\delta = 1$. Using confidence level $1 - \alpha = 0.9$ and $n_0 = 10$ initial replications for the $k + 1 = 7$ policies, the critical values for Case D are $h = 2.510$ and $g = 3.951$; Case D critical values were employed because we simulated all systems using CRN. Because smaller expected cost is better, an alternative must have $c = h\delta/g = \$0.64$ lower sample mean cost than the FIFO policy in order to replace it.

Let $S_{i\ell}^2$ denote the sample variance of the difference between systems i and ℓ. The first-stage experiment gave the sample variances shown in Table 6. The dramatic effect of CRN can be seen by looking at Table 7, which shows the marginal sample variances S_i^2 of each system. If CRN were not employed, or if a procedure were used that did not account for CRN, then the variance of the difference between systems i and ℓ would be approximately the sum $S_i^2 + S_\ell^2$.

Table 6 **Sample Variances of the Differences, $S^2_{i\ell}$, Based on the Initial $n_0 = 10$ Observations**

	SEIZE1	SEIZE2	TH1(3)	TH2(3)	PREEMPT1	PREEMPT2
FIFO	0.62	3.27	0.43	0.36	3.80	0.71
SEIZE1		6.60	2.01	0.05	7.36	0.02
SEIZE2			1.45	5.74	0.04	6.91
TH1(3)				1.54	1.82	2.17
TH2(3)					6.45	0.07
PREEMPT1						7.68

Since $\widehat{S}^2 = \max_{i,\,\ell:i\neq\ell} S^2_{i\ell} = 7.682$, the second-stage sample size is

$$N = \max\left\{10, \left\lceil \frac{(3.951)^2(7.682)}{1^2} \right\rceil\right\} = 100.$$

Therefore, the total sample size from all systems is 700 observations, giving the second-stage sample means shown in Table 8.

Since $\overline{X}_6 = 4.69 < \overline{X}_0 - 0.64 = 5.15$, we conclude that PREEMPT2 is better (has smaller expected cost) than the standard, FIFO. But we can say more. Table 8 also displays the 90% MCB upper and lower confidence bounds (UCB and LCB, respectively). These indicate the FIFO, SEIZE2, TH1(3), and PREEMPT1 can with high confidence be declared inferior to the best, because the LCB on $\mu_i - \min_{\ell\neq i}\mu_\ell$ is 0. However, SEIZE1 and TH2(3) might in fact be the best, because their confidence intervals for difference from the best contain 0. We are assured (with 90% confidence) that even if PREEMPT2 is not the best, its expected cost is within \$0.79 of the least expected cost.

To compare these results to an existing procedure that does not account for CRN, we applied Dudewicz's and Dalal's (1983) two-stage procedure for forming two-sided, fixed-width-δ confidence intervals for $\mu_i - \mu_0$, $i = 1, 2, \ldots, k$ (another similar option is the procedure of Bofinger and Lewis 1992). This procedure could be used to discover all policies whose expected cost is more than $\delta = \$1$ different from FIFO's expected cost. Their procedure only makes use of the marginal variances in Table 7. When applying the procedure to this example we found that 981 observations were required—281 more than our procedure. Further, this procedure provides inference on the difference between each policy and FIFO, but not direct inference about the best policy.

Table 7 **Marginal Sample Variances, S^2_i, Based on the Initial $n_0 = 10$ Observations**

FIFO	SEIZE1	SEIZE2	TH1(3)	TH2(3)	PREEMPT1	PREEMPT2
6.87	3.48	19.57	10.60	4.12	20.82	3.26

Table 8 **Second-Stage Sample Means, $\overline{X}_i$, Based on 100 Observations**

i	0	1	2	3	4	5	6
policy	FIFO	SEIZE1	SEIZE2	TH1(3)	TH2(3)	PREEMPT1	PREEMPT2
$\overline{X}_i$	5.79	4.90	7.79	6.29	5.36	8.19	4.69
$\overline{X}_i - \min_{\ell\neq i}\overline{X}_\ell$	1.10	0.21	3.10	1.60	0.67	3.50	−0.21
MCB UCB	2.10	1.21	4.10	2.60	1.67	4.50	0.79
MCB LCB	0	−0.79	0	0	−0.33	0	−1.21

7. Extensions

We have presented procedures for comparisons with a standard that generalize procedures due to Bechhofer and Turnbull (1978) and Paulson (1952) so that they are more useful for the types of data encountered in simulation experiments. Our procedures also add confidence bounds on certain differences, bounds that were not present in the earlier procedures.

While we do generalize the procedures of Bechhofer and Turnbull (1978) with respect to the types of data to which they apply, Bechhofer and Turnbull consider a more general decision problem than ours. Specifically, they derive procedures that guarantee that

$$\Pr\{\text{select } \pi_0\} \geq 1-\alpha_0 \text{ whenever } \mu_0 \geq \mu_{[k]} + \delta_0 \text{ and}$$

$$\Pr\{\text{select } \pi_{[k]}\} \geq 1-\alpha_1 \text{ whenever } \mu_{[k]} \geq \mu_{[k-1]} + \delta_1, \text{ and } \mu_{[k]} \geq \mu_0 + \delta_2.$$

Notice that this formulation allows for three indifference-zone parameters and a different probability requirement for retaining the standard versus selecting the best alternative. Our procedures can all be extended to cover this more general case.

Acknowledgments

This research was partially supported by National Science Foundation Grant numbers DMI-9622065 and DMI-9622269, and by Rockwell Software and Symix Corporation. The authors gratefully acknowledge the assistance provided by the associate editor and three referees.

References

Bechhofer, R. E., B. W. Turnbull. 1978. Two $(k+1)$-decision selection procedures for comparing k normal means with a specified standard. *J. Amer. Statist. Assoc.* **73** 385–392.

Bofinger, E., G. J. Lewis. 1992. Two stage procedures for multiple comparisons with a control. E.J. Dudewicz, eds. *The Frontiers of Modern Statistical Inference Procedures*, Vol. II. American Sciences Press, Inc,

Desruelle, P., H. J. Steudel. 1996. A queuing network model of a single operator manufacturing workcell with machine/operator interference. *Management Sci.* **42** 576–590.

Dudewicz, E. J., S. R. Dalal. 1983. Multiple-comparisons with a control when variances are unknown and unequal. *Amer. J. Math. Management Sci.* **3** 275–295.

Goldsman, D., B. L. Nelson, B. Schmeiser. 1991. Methods for selecting the best system. B. L. Nelson, W. D. Kelton, G. M. Clark, eds., *Proceedings of the 1991 Winter Simulation Conference*. Institute of Electrical and Electronics Engineers, Piscataway, NJ. 177–186.

Hsu, J. 1996. *Multiple Comparisons*. Chapman & Hall, London, U.K.

Marsaglia, G., I. Olkin. 1984. Generating correlation matrices. *SIAM J. Sci. Statist. Comput.* **5** 470–475.

Matejcik, F. J., B. L. Nelson. 1995. Two-stage multiple comparisons with the best for computer simulation. *Oper. Res.* **43** 633–640.

Nakade, K., K. Ohno, J. G. Shanthikumar. 1997. Bounds and approximations for cycle time of a U-shaped production line. *Oper. Res. Lett.* **21** 191–200.

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

Nelson, B. L. 1993. Robust multiple comparisons under common random numbers. *ACM Trans. Modeling Comput. Simulation* **3** 225–243.

———, S. Banerjee. 2001. Selecting a good system: Procedures and inference. *IIE Trans.*, Forthcoming.

———, F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Sci.* **41** 1935–1945.

Paulson, E. 1952. On the comparison of several experimental categories with a control. *Ann. Math. Statist.* **23** 239–246.

Tong, Y. L. 1980. *Probability Inequalities in Multivariate Distributions.* Academic Press, New York.

Yang, W., B. L. Nelson. 1991. Using common random numbers and control variates in multiple-comparison procedures. *Oper. Res.* **39** 583–591.

Accepted by Pierre L'Ecuyer; received June 19, 1997. This paper was with the authors 16 months for 3 revisions.