Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

SELECTING THE BEST SYSTEM: THEORY AND METHODS

Seong-Hee Kim

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.

Barry L. Nelson

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

ABSTRACT

This paper provides an advanced tutorial on the construction of ranking-and-selection procedures for selecting the best simulated system. We emphasize procedures that provide a guaranteed probability of correct selection, and the key theoretical results that are used to derive them.

1 INTRODUCTION

Over the last twenty years there has been considerable effort expended to develop statistically valid ranking-and-selection (R&S) procedures to compare a finite number of simulated alternatives. There exist at least four classes of comparison problems that arise in simulation studies: selecting the system with the largest or smallest expected performance measure (selection of the best), comparing all alternatives against a standard (comparison with a standard), selecting the system with the largest probability of actually being the best performer (multinomial selection), and selecting the system with the largest probability of success (Bernoulli selection). For all of these problems, a constraint is imposed either on the probability of correct selection (PCS) or on the simulation budget. Some procedures find a desirable system with a guarantee on the PCS, while other procedures maximize the PCS under the budget constraint. Our focus is on selection-of-the-best problems with a PCS constraint. A good procedure is one that delivers the desired PCS efficiently (with minimal simulated data) and is robust to modest violations of its underlying assumptions. Other types of comparison problems and procedures will be discussed briefly in Section 7. In this tutorial "best" means maximum expected value of performance, such as expected throughput or profit.

Rather than present a comprehensive survey of R&S procedures, or provide a guide for applying them, our goal is to explain how such procedures are constructed, emphasizing issues that are central to designing procedures for computer simulation, and reviewing some key theorems that have proven useful in deriving procedures. We do, however, present two specific R&S procedures as illustrations. See Goldsman and Nelson (1998) and Law and Kelton (2000) for detailed "how to" guides, and Bechhofer et al. (1995) for a comprehensive survey of R&S procedures.

The paper is organized as follows: In Section 2 we show how R&S procedures are derived in an easy, but unrealistic, setting. Section 3 lists the challenges and opportunities encountered in simulation problems, along with key theorems and results that have proven useful in extending R&S procedures to this setting. Two specific procedures are presented in Section 4, followed by a numerical illustration in Section 5. Section 6 briefly reviews asymptotic analysis regimes for R&S. Finally, Section 7 closes the paper by describing other formulations of the R&S problem and giving appropriate references.

2 BASICS OF RANKING AND SELECTION

In this section we employ the simplest possible setting to illustrate how R&S procedures attack comparison problems. This setting (i.i.d. normal data with known, common variance) allows us to focus on key techniques before moving on to the technical difficulties that arise in designing procedures for realistic simulation problems.

R&S traces its origins to two papers: Bechhofer (1954) established the *indifference-zone formulation*, while Gupta (1956, 1965) is credited with the *subset selection formulation* of the problem. Both approaches are reviewed in this section, and both were developed to compensate for the limited inference provided by hypothesis tests for the homogeniety of the k population parameters (usually means). In many industrial and biostatistics experiments, rejecting the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, where μ_i is the parameter associated with the kth population, leads naturally to questions about which one has the largest or smallest parameter. R&S tries to answer such questions. Multiple comparison procedures (MCPs) also provide inference beyond rejection of homogeniety; there is a close connection between R&S and MCPs, as we demonstrate later.

Suppose that there are k systems. Let X_{ij} represent the jth i.i.d. output from system i and let $\mathbf{X}_i = \{X_{ij} ; j = 1, 2, \ldots\}$ denote the output sequence from system i. In this section, we assume that the X_{ij} are normally distributed with means $\mu_i = E[X_{ij}]$ and variances $\sigma_i^2 = \text{Var}[X_{ij}]$. Further, we assume that the processes $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ are mutually independent, and the variances are known and equal; that is, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. These are unrealistic assumptions that will be relaxed later, but we adopt them here because we can derive R&S procedures in a way that illustrates the key issues. Throughout the paper we assume that a larger mean is better, and we let $\mu_k \geq \mu_{k-1} \geq \cdots \geq \mu_1$, so that (unknown to us) system k is the best system.

2.1 Subset-Selection Formulation

Suppose that we have n outputs from each of the systems. Our goal is to use this data to obtain a subset $I \subseteq \{1, 2, \ldots, k\}$ such that

$$\Pr\{k \in I\} \geq 1 - \alpha \tag{1}$$

where $1/k < 1 - \alpha < 1$. Ideally $|I|$ is small, the best case being $|I| = 1$. Gupta's solution was to include in the set I all systems ℓ such that

$$\bar{X}_\ell(n) \geq \max_{i \neq \ell} \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}} \tag{2}$$

where $\bar{X}_i(n)$ is the sample mean of the (first) n outputs from system i, and h is a constant whose value will depend on k and $1 - \alpha$. The proof that rule (2) provides guarantee (1) is instructive and shows what the value of h should be:

$$\begin{aligned}
&\Pr\{k \in I\} \\
&\quad = \Pr\left\{\bar{X}_k(n) \geq \max_{i \neq k} \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}}\right\} \\
&\quad = \Pr\left\{\bar{X}_k(n) \geq \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}}, \forall i \neq k\right\} \\
&\quad = \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} \right. \\
&\qquad\qquad \left. \leq h - \frac{(\mu_i - \mu_k)}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\quad \geq \Pr\{Z_i \leq h, i = 1, 2, \ldots, k-1\} = 1 - \alpha
\end{aligned}$$

where $(Z_1, Z_2, \ldots, Z_{k-1})$ have a multivariate normal distribution with means 0, variances 1, and common pairwise correlations $1/2$. Therefore, to provide the guarantee (1), h needs to be the $1 - \alpha$ quantile of the maximum of such a multivariate normal random vector, a quantile that turns out to be relatively easy to approximate numerically. Notice the inequality in the final step where we make use of the fact that $\mu_k \geq \mu_i$.

A theme that runs throughout much of R&S is first using appropriate standardization of estimators and then bounding the resulting probability statements in such a way that a difficult multivariate probability statement becomes one that is readily solvable.

2.2 Indifference-Zone Formulation

A disadvantage of the subset-selection procedure in Section 2.1 is that the retained set I may, and likely will, contain more than one system. However, there is no procedure that can guarantee a subset of size 1 and satisfy (1) for arbitrary n. Even when n is under our control, as it is in computer simulation, the appropriate value will depend on the true differences $\mu_k - \mu_i, \forall i \neq k$. To attack this problem, Bechhofer suggested the following compromise: guarantee to select the single best system, k, whenever $\mu_k - \mu_{k-1} \geq \delta$, where $\delta > 0$ is the smallest difference the experimenter feels is worth detecting. Specifically, the procedure should guarantee

$$\Pr\{\text{select } k | \mu_k - \mu_{k-1} \geq \delta\} \geq 1 - \alpha \tag{3}$$

where $1/k < 1 - \alpha < 1$. If there are systems whose means are within δ of the best, then the experimenter is "indifferent" to which of these is selected, leading to the term indifference-zone (IZ) formulation.

The procedure is as follows: From each system, take

$$n = \left\lceil \frac{2h^2\sigma^2}{\delta^2} \right\rceil \tag{4}$$

outputs, where h is an appropriate constant (determined below) and $\lceil x \rceil$ means to round x up; then select the system with the largest sample mean as the best. Assuming $\mu_k - \mu_{k-1} \geq \delta$,

$$\begin{aligned}
&\Pr\{\text{select } k\} \\
&\quad = \Pr\left\{\bar{X}_k(n) > \bar{X}_i(n), \forall i \neq k\right\} \\
&\quad = \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < -\frac{(\mu_i - \mu_k)}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\quad \geq \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < \frac{\delta}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\quad \geq \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < h, \forall i \neq k\right\} \\
&\quad = \Pr\{Z_i < h, i = 1, 2, \ldots, k-1\} = 1 - \alpha
\end{aligned}$$

where again $(Z_1, Z_2, \ldots, Z_{k-1})$ has a multivariate normal distribution with means 0, variances 1, and common pairwise

correlations $1/2$, implying h needs to be the $1-\alpha$ quantile of the maximum of such a multivariate normal random vector.

Notice that the first inequality results from the assumption that $\mu_k - \mu_{k-1} \geq \delta$, while the second occurs because $\sqrt{n} \geq \sqrt{2}h\sigma/\delta$. Both of these tricks are standard: the first frees the probability statement of dependence on the true means, while the second frees it of dependence on the value of the variance.

It is worth noting that, over all configurations of the true means such that $\mu_k - \mu_{k-1} \geq \delta$, the configuration $\mu_i = \mu_k - \delta, \forall i \neq k$ minimizes the PCS; it is therefore known as the *least-favorable configuration* (LFC). In this paper we break from the statistics literature in that we will not be concerned with identifying the LFC; our only interest is insuring that (3) is met.

Bechhofer's procedure is essentially a power calculation: how large a sample is required to detect differences of at least δ? When true differences are greater than δ, Bechhofer's n may be much larger than needed. By taking observations and making decisions sequentially, it is often possible to reach an earlier decision. Sequential selection procedures can be traced back at least to Wald (1947), but here we present a procedure due to Paulson (1964) that better illustrates the approach that has had the most impact in computer simulation. Paulson's procedure takes observations *fully sequentially*—meaning one at a time—and *eliminates* systems from continued sampling when it is statistically clear that they are inferior. Thus, a problem with a single dominant alternative may terminate very quickly.

Using the same notation as above, let $\bar{X}_i(r)$ be the sample mean of the first r outputs of system i. At each stage $r = 1, 2, \ldots, n$, one output is taken from each system whose index is in I, where initially $I = \{1, 2, \ldots, k\}$. At stage r, system ℓ is retained in I only if

$$\bar{X}_\ell(r) \geq \max_{i \in I} \bar{X}_i(r) - \max\{0, a/r - \lambda\} \tag{5}$$

where $a > 0$ and $0 < \lambda < \delta$ are constants to be determined, and $n = \lfloor a/\lambda \rfloor$, with $\lfloor \cdot \rfloor$ meaning round down. The procedure ends when $|I| = 1$, which requires no more than $n + 1$ stages. Parallels with Gupta's subset selection and Bechhofer's IZ ranking are obvious: At each stage a subset selection is performed, with the hedging factor $(a/r - \lambda)$ decreasing as more data are obtained. In the end, if the procedure makes it that far, the system with the largest sample mean is selected.

The following result is used to establish the PCS: Suppose $Z_1, Z_2, \ldots$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$ with $\mu < 0$. Then it can be shown that

$$\Pr\left\{\bar{Z}(r) > \frac{a}{r}, \text{ for some } r < \infty\right\} \leq \exp\left(\frac{2\mu}{\sigma^2}a\right). \tag{6}$$

Large deviation results, frequently based on the analysis of approximating Brownian motion processes, are central to the design of fully sequential procedures that involve frequent looks at the data.

The approach in this case is to bound the probability of an *incorrect selection* (ICS). An ICS event occurs if system k is eliminated at some point during the procedure. Let $\Pr\{\mathrm{ICS}_i\}$ be the probability of an incorrect selection if only systems i and k are included in the competition.

The first key inequality is

$$\Pr\{\mathrm{ICS}\} \leq \sum_{i=1}^{k-1} \Pr\{\mathrm{ICS}_i\}. \tag{7}$$

Decomposition into some form of paired comparisons is a key step in many sequential procedures.

This decomposition allows us to focus only on $\Pr\{\mathrm{ICS}_i\}$. Notice that

$$\begin{aligned}
&\Pr\{\mathrm{ICS}_i\} \\
&\quad\leq \Pr\left\{\bar{X}_k(r) < \bar{X}_i(r) + \lambda - a/r, \text{ for some } r \leq n+1\right\} \\
&\quad= \Pr\left\{\bar{X}_i(r) - \bar{X}_k(r) + \lambda > a/r, \text{ for some } r \leq n+1\right\} \\
&\quad\leq \Pr\left\{\bar{X}_i(r) - \bar{X}_k(r) + \lambda > a/r, \text{ for some } r < \infty\right\} \\
&\quad\leq \exp\left(\frac{(\mu_i - \mu_k + \lambda)}{\sigma^2}a\right) \\
&\quad\leq \exp\left(\frac{(\lambda - \delta)}{\sigma^2}a\right).
\end{aligned}$$

The third inequality comes from the large deviation result (6), while the fourth inequality exploits the indifference-zone assumption. If we set

$$a = \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{\delta - \lambda} \tag{8}$$

then $\Pr\{\mathrm{ICS}_i\} \leq \alpha/(k-1)$ and

$$\Pr\{\mathrm{ICS}\} \leq (k-1)\frac{\alpha}{(k-1)} = \alpha.$$

2.3 Connection to Multiple Comparisons

MCPs approach the comparison problem by providing simultaneous confidence intervals on selected differences among the systems' parameters. Hochberg and Tamhane (1987) and Hsu (1996) are good comprehensive references. As noted by Hsu (1996, pp. 100-102), the connection between R&S and MCPs comes through multiple comparisons with the best (MCB). MCB forms simultaneous confidence intervals for $\mu_i - \max_{\ell \neq i} \mu_\ell, i = 1, 2, \ldots, k$, the difference

between each system and the best of the rest. Specialized to the known-variance case, the intervals take the form

$$\mu_i - \max_{\ell \neq i} \mu_\ell \in \left[-\left(\bar{X}_i(n) - \max_{\ell \neq i} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}} \right)^-, \left(\bar{X}_i(n) - \max_{\ell \neq i} \bar{X}_\ell(n) + h\sigma\sqrt{\frac{2}{n}} \right)^+ \right] \quad (9)$$

where h is the same critical value used in Bechhofer's and Gupta's procedures, $-x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$. Under our assumptions these k confidence intervals are simultaneously correct with probability $\geq 1 - \alpha$.

Consider the set I containing the indices of all systems whose MCB upper confidence bound is greater than 0. Thus, for $i \in I$,

$$\bar{X}_i(n) > \max_{\ell \neq i} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}}$$

meaning these are the same systems that would be retained by Gupta's subset-selection procedure. Since $\mu_k - \max_{\ell \neq k} \mu_\ell > 0$, and these intervals are simultaneously correct with probability $\geq 1 - \alpha$, system k will be in the subset identified by the MCB upper bounds with the required probability.

Now suppose that n has been selected such that $n \geq 2h^2\sigma^2/\delta^2$, implying that

$$h\sigma\sqrt{\frac{2}{n}} \leq \delta$$

as in Bechhofer's procedure. Let B be the index of the system with the largest sample mean. Then the MCB lower bounds guarantee with probability $\geq 1 - \alpha$ that

$$\begin{aligned} \mu_B - \max_{\ell \neq B} \mu_\ell &\geq -\left(\bar{X}_B(n) - \max_{\ell \neq B} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}} \right)^- \\ &\geq -\delta. \end{aligned}$$

The final inequality follows because $\bar{X}_B(n) - \max_{\ell \neq B} \bar{X}_\ell(n) \geq 0$ by the definition of B, and $h\sigma\sqrt{2/n} \leq \delta$ because of our choice of n. As noted by Nelson and Goldsman (2001), this establishes that the system selected by Bechhofer's procedure is guaranteed to be within δ of the true best *under any configuration of the means*. Further, if $\mu_k - \mu_{k-1} > \delta$, then $\Pr\{B = k\} \geq 1 - \alpha$ as required.

As a consequence of this analysis both Bechhofer's and Gupta's procedures can be augmented with MCB confidence intervals "for free," and Bechhofer's procedure is guaranteed to select a system within δ of the best. Nelson and Matejcik (1995) establish very mild conditions under which these results hold for far more general R&S procedures.

3 SIMULATION ISSUES AND KEY RESULTS

In the previous section we illustrated different approaches to the R&S problem under assumptions such as independence, normality, and known and equal variances. Unfortunately, such assumptions rarely hold in simulation experiments. There are also opportunities available in simulation experiments that are not present in physical experiments. In the following subsections we describe these issues and opportunities, and present key theorems and results that have been useful in deriving R&S procedures that overcome or exploit them.

3.1 Unknown and Unequal Variances

Unknown and unequal variances across alternatives is a fact of life in system simulation problems, and the variances can differ dramatically. In the simple inventory model presented in Section 5 the ratio of the largest to smallest variance is almost 4.

There are many subset-selection procedures that permit an unknown, common variance (see Goldsman and Nelson 1998 for one). When variances are unknown and unequal, however, the subset-selection problem is essentially equivalent to the famous Behrens-Fisher problem. One approach is to work with the standardized random variables

$$\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\left(\frac{S_i^2}{n} + \frac{S_k^2}{n} \right)^{1/2}}, i = 1, 2, \ldots, k-1. \quad (10)$$

Neither the joint nor marginal distributions of these quantities are conveniently characterized. If you break the required joint probability statement up into statements about the individual terms, using techniques described below, then there are at least two solutions available. Welch (1938) suggested approximating each term in (10) as having a $t_{\hat{\nu}}$ distribution, where the degrees of freedom $\hat{\nu}$ is an approximation based on the values of S_i^2 and S_k^2. Banerjee (1961) proposed a probability bound that we specialize to our case:

Theorem 1 (Banerjee 1961) *Suppose Z is* $\mathrm{N}(0, 1)$ *and independent of Y_i and Y_k, which are themselves independent χ^2_ν random variables. Then for arbitrary but fixed* $0 \leq \gamma \leq 1$,

$$\Pr\left\{ \frac{Z^2}{\gamma \frac{Y_i}{\nu} + (1 - \gamma)\frac{Y_k}{\nu}} \leq t^2_{1-\alpha/2,\nu} \right\} \geq 1 - \alpha. \quad (11)$$

To employ Banerjee's inequality in our context, identify

$$Z = \frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\left(\frac{\sigma_i^2}{n} + \frac{\sigma_k^2}{n}\right)^{1/2}}$$

and

$$\gamma\frac{Y_i}{\nu} + (1-\gamma)\frac{Y_k}{\nu} = \frac{\frac{S_i^2}{n} + \frac{S_k^2}{n}}{\frac{\sigma_i^2}{n} + \frac{\sigma_k^2}{n}}$$

$$= \left(\frac{\sigma_i^2}{\sigma_i^2 + \sigma_k^2}\right)\frac{S_i^2}{\sigma_i^2} + \left(\frac{\sigma_k^2}{\sigma_i^2 + \sigma_k^2}\right)\frac{S_k^2}{\sigma_k^2}.$$

This inequality is used in Procedure NSGS presented in Section 4.

For some time it has been known that it is not possible to provide a guaranteed PCS, in the IZ sense, with a single stage of sampling when variances are unknown (see Dudewizc 1995 for a comprehensive discussion of this result). Thus, practically useful IZ procedures work sequentially—meaning two or more stages of sampling—with the first stage providing variance estimates that help determine how much, if any, additional sampling is needed in the succeeding stages. However, one cannot simply substitute variance estimators into Bechhofer's or Paulson's procedures and hope to achieve a guaranteed PCS. Instead, the uncertainty in the variance estimators enters into the determination of the sample sizes, invariably leading to more sampling than would take place if the variances were known.

A fundamental result in parametric statistics is the following: If $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent random variables. The result extends in the natural way to random vectors $\mathbf{X}_j$ that are multivariate normal. An extension of a different sort, due to Stein (1945), is fundamental to R&S procedures with unknown variances:

Theorem 2 (Stein 1945) *Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, and S^2 is $\sigma^2\chi_\nu^2/\nu$ and independent of $\sum_{i=1}^n X_j$ and of $X_{n+1}, X_{n+2}, \ldots$.*

1. *If $N \geq n$ is a function only of S^2 then*

$$\frac{\bar{X}(N) - \mu}{S/\sqrt{N}} \sim t_\nu. \tag{12}$$

2. *If $\xi > 0$ and*

$$N = \max\left\{\left\lceil \frac{S^2}{\xi^2} \right\rceil, n+1\right\}$$

then for any weights $w_1, w_2, \ldots, w_N$ satisfying $\sum_{j=1}^N w_j = 1$, $w_1 = w_2 = \cdots = w_n$, and $S^2\sum_{j=1}^N w_j^2 = \xi^2$ we have

$$\frac{\sum_{j=1}^N w_j X_j - \mu}{\xi} \sim t_\nu. \tag{13}$$

In the usual case where S^2 is the sample variance of the first n observations, $\nu = n-1$. The importance of this result in R&S is that it allows determination of a sample size large enough to attain the desired power against differences of at least δ without requiring knowledge of the process variance.

If comparisons of only $k = 2$ systems were necessary, then Stein's result would be enough (at least in the i.i.d. normal case). But our problem is multivariate, making joint probability statements about

$$\frac{\bar{X}_i(N_i) - \bar{X}_k(N_k) - (\mu_i - \mu_k)}{\mathcal{S}_{ik}}, \; i = 1, 2, \ldots, k-1 \tag{14}$$

where $\mathcal{S}_{ik}^2$ is a variance estimate based on an initial sample of size (say) n, and N_i and N_k are the final sample sizes from systems i and k. The joint distribution of these random variables is quite messy in general, even if all systems are simulated independently (as we assume in this section). One approach is to condition on $\mathcal{S}_{ik}$ and $\bar{X}_k(N_k)$ and apply inequalities such as the following to bound the joint probability:

Theorem 3 (Kimball 1951) *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*

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

Kimball's theorem is actually only the case $k = 1$; see Hochberg and Tamhane (1987) for the extension.

Theorem 4 (Slepian 1962) *Let $(Z_1, Z_2, \ldots, Z_k)$ have a k-variate normal distribution with zero mean vector, unit variances, and correlation matrix $\mathbf{R} = \{\rho_{ij}\}$. Let $\xi_1, \xi_2, \ldots, \xi_k$ be some constants. If all the $\rho_{ij} \geq 0$, then*

$$\Pr\left\{\bigcap_{i=1}^k (Z_i \leq \xi_i)\right\} \geq \prod_{i=1}^k \Pr\{Z_i \leq \xi_i\}.$$

Notice that, conditional on the $\mathcal{S}_{ik}^2$, the terms in (14) are positively correlated (due to the common $\bar{X}_k(N_k)$ term), providing the opening to apply Slepian's inequality. Kimball's inequality then can be applied to simplify the uncondition-

ing on S_{ik}^2. Both of these ideas are employed in the design of Procedure NSGS below.

3.2 Non-Normality of Output Data

Raw output data from industrial and service simulations are rarely normally distributed. Surprisingly, non-normality is usually not a concern in simulation experiments that (a) are designed to make multiple independent replications, and (b) use a within-replication average of a large number of raw simulation outputs as the basic summary measure. This is frequently the situation for so-called "terminating simulations" in which the initial conditions and stopping time for each replication are an inherent part of the definition of the system. A standard example is a store that opens empty at 6 AM, then closes when the last customer to arrive before 9 PM leaves the store. If the output of interest is the average customer delay in the checkout line over the course of the day, and comparisons will be based on the expected value of this average, and the average is over many individual customer delays, then the Central Limit Theorem suggests that the replication averages will be approximately normally distributed.

Difficulties arise in so-called "steady-state simulations" where the parameter of interest is defined by a limit as the time index of a stochastic process approaches infinity (and therefore forgets its initial conditions). Some steady-state simulations are amenable to multiple replications of each alternative and within-replication averages as summary statistics, in which case the preceding discussion applies. Unfortunately, severe estimator bias due to residual effects of the initial conditions sometimes force an experiment design consisting of a single, long replication from each alternative. The raw outputs within each replication are typically neither normally distributed nor independent. For example, waiting times of individual customers in a queueing system are usually dependent because a long delay for one customer tends to increase the delays of the customers who follow. The best we can hope for is an approximately stationary output process from each system, but not normality or independence.

The most common approach for dealing with this problem is to transform the raw data into *batch means*, which are averages of large number of raw outputs. The batch means are often far less dependent and non-normal than the raw output data. There are problems with the batching approach for R&S, however. If a "stage" is defined by batch means rather than raw output, then the simulation effort consumed by a stage is a multiple of the batch size. When a large batch size is required to achieve approximate independence—and batch sizes of several thousand are common—then the selection procedure is forced to make decisions at long intervals, wasting outputs and time. This inefficiency becomes serious when fully sequential procedures are employed because the elimination decisions for clearly inferior systems must wait for an entire batch to be formed. Therefore, for steady-state simulations, selection procedures that use individual raw outputs as basic observations are desirable.

Although no known procedures provide a guaranteed PCS for single-replication designs, some procedures have shown good empirical performance (e.g., Sullivan and Wilson 1989), while others have been shown to be asymptotically valid. See Law and Kelton (2000) for a general discussion of replications versus batching, Glynn and Iglehart (1990) for conditions under which the batch means method is asymptotically valid for confidence intervals, and Section 6 for a review of asymptotic analysis of R&S procedures.

3.3 Common Random Numbers

The procedures described in Section 2 assumed that data across the k alternative systems are independent. In simulation experiments this assumption can be made valid by using different sequences of random numbers to drive the simulation of each system. However, since we are making comparisons, there is a potential advantage to using common random numbers (CRN) to drive the simulation of each system because

$$\mathrm{Var}[X_{ij} - X_{\ell j}] = \mathrm{Var}[X_{ij}] + \mathrm{Var}[X_{\ell j}] - 2\mathrm{Cov}[X_{ij}, X_{\ell j}].$$

If implemented correctly (see, for instance, Banks, et al. 2001), CRN tends to make $\mathrm{Cov}[X_{ij}, X_{\ell j}] > 0$ thereby reducing the variance of the difference.

R&S procedures often need to make probability statements about the collection of random variables

$$\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k), i = 1, 2, \ldots, k-1. \quad (15)$$

The appearance of the common term $\bar{X}_k(n)$ causes dependence among these random variables, but it is often easy to model or tightly bound. The introduction of CRN induces dependence between $\bar{X}_i(n)$ and $\bar{X}_k(n)$ as well. Even though the sign of the induced covariance is believed known, its value is not, making it difficult to say anything about the dependence among the differences (15).

Two approaches are frequently used. The first is to replace the basic data $\{X_{ij}; i = 1, 2, \ldots, k; j = 1, 2, \ldots, n\}$ with pairwise differences $\{X_{ij} - X_{\ell j}; i \neq \ell; j = 1, 2, \ldots, n\}$ because the variance of the sample mean of the difference includes the effect of the CRN-induced covariance. The second is to apply the Bonferroni inequality to break up joint statements about (15) into statements about the individual terms. Recall that for events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{k-1}$, the Bonferroni inequality states that

$$\Pr\left\{\bigcap_{i=1}^{k-1} \mathcal{E}_i\right\} \geq 1 - \sum_{i=1}^{k-1} \Pr\left\{\mathcal{E}_i^c\right\}. \quad (16)$$

In the R&S context $\mathcal{E}_i$ might correspond to an event like $\{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k) \leq h\}$.

Approaches based on the Bonferroni inequality make no assumption about the induced dependence, and therefore are very conservative. A more aggressive approach is to assume some structure for the dependence induced by CRN. One standard assumption is that all pairwise correlations $\rho = \text{Corr}[X_{ij}, X_{\ell j}]$ are positive, but identical, and all variances are equal; this is known as *compound symmetry*. Nelson and Matejcik (1995) extended Rinott's procedure (1978)—one of the simplest and most popular IZ procedures—in conjunction with CRN under a more general structure called *sphericity*. The specific assumption is

$$\text{Cov}[X_{ij}, X_{\ell j}] = \begin{cases} 2\beta_i + \tau^2, & i = \ell \\ \beta_i + \beta_\ell, & i \neq \ell \end{cases} \tag{17}$$

with $\tau^2 > 0$, which is equivalent to assuming that $\text{Var}[X_{ij} - X_{\ell j}] = 2\tau^2$ for all $i \neq \ell$, a type of variance balance. This particular structure is useful because there exists an estimator $\widehat{\tau}^2$ of τ^2 that is independent of the sample means and has a χ^2 distribution (allowing a pivotal quantity to be formed and Stein's theorem to be applied). Nelson and Matejcik (1995) showed that procedures based on this assumption are robust to departures from sphericity, at least in part because assuming sphericity is like assuming that all pairwise correlations equal the average pairwise correlation.

3.4 The Sequential Nature of Simulation

Suppose an IZ ranking procedure is applied in the study of k new blood pressure medications. Then "replications" correspond to patients, and the idea of using a fully sequential procedure (assign one patient at a time to each drug, then wait for the results before recruiting the next patient) seems absurd. In simulation experiments, however, data are naturally generated sequentially, at least within each simulated alternative, making multi-stage procedures much more attractive. However, there are some issues:

- In multiple-replication designs, sequential sampling is particularly attractive. All that needs to be retained to start the next stage of sampling is the ending random number seeds from the previous stage. In single-replication designs it can be more difficult to resume sampling from a previous stages, since the entire state of the system must be retained and restored.
- A hidden cost of using multi-stage procedures is the computational overhead in switching among the simulations of the k alternatives. On a single-processor computer, switching can involve saving output, state and seed information from the current system; swapping the program for the current system out of, and for the next system into, active memory; and restoring previous state and seed information for the next system. Thus, the overall computation effort includes both the cost of generating simulated data and the cost of switching. Hong and Nelson (2003) look at sequential IZ procedures that attempt to minimize the total computational cost.
- If k processors are available, then an attractive option is to assign each system to a processor and simulate in parallel. This is highly effective in conjunction with R&S procedures that require little or no coordination between the simulations of each system, such as subset-selection procedures or IZ-ranking procedures that use only variance information (and not differences among the sample means). Unfortunately, a fully sequential procedure with elimination would defeat much of the benefit of parallel processing because communication among the processors is required after generating each output.

Many sequential procedures are based on results for Brownian motion processes. Let $\mathcal{B}(t; \Delta)$ be a standard Brownian motion process with drift Δ. Consider the partial sum of the pairwise difference $D_i(r) = \sum_{j=1}^{r}(X_{kj} - X_{ij})$, $r = 1, 2, \ldots$. If the X_{ij} are i.i.d. normal, and $\mu_k - \mu_i = \delta$, then $\{D_i(r), r = 1, 2, \ldots\} \stackrel{\mathcal{D}}{=} \{\sigma\mathcal{B}(t; \delta/\sigma), t = 1, 2, \ldots\}$, where $\sigma^2 = \text{Var}[X_{kj} - X_{ij}]$ (with or without CRN). In other words, $D_i(r)$ is a Brownian motion process with drift observed only at discrete (integer) points in time. A great deal is known about the probability of Brownian motion processes crossing boundaries in various ways (see, for instance, Siegmund 1985 or Jennison and Turnbull 2000); we display one specific result below. Thus, it seems natural to design R&S procedures for $\sigma\mathcal{B}(t; \delta/\sigma)$ and apply them to $D_i(r)$.

Let $c(t)$ be a symmetric (about 0) continuation region for $\sigma\mathcal{B}(t; \delta/\sigma)$, and let an incorrect selection correspond to the process exiting the region in the wrong direction (down, when the drift is positive). If $T = \inf\{t \geq 0 : |\sigma\mathcal{B}(t; \delta/\sigma)| > c(t)\}$, then

$$\Pr\{\text{ICS}_i\} = \Pr\{\sigma\mathcal{B}(T; \delta/\sigma) < 0\}.$$

Of course $\sigma\mathcal{B}(t; \delta/\sigma)$ is only an approximation for $D_i(r)$. However, Jennison, et al. (1980) show that under very general conditions, $\Pr\{\text{ICS}_i\}$ is no greater if the Brownian motion process is observed at discrete times; thus, procedures designed for $\sigma\mathcal{B}(t; \delta/\sigma)$ provide an upper bound on the probability of incorrect selection for $D_i(r)$. In conjunction with a decomposition into pairwise comparisons, as in (7), this result can be used to derive R&S procedures for $k \geq 2$.

Fabian (1974) tightened the triangular continuation region used by Paulson, and this was exploited by Hartmann

(1988, 1991), Kim and Nelson (2001, 2003) and Hong and Nelson (2003).

Theorem 5 (Fabian 1974) *Let $\{\mathcal{B}(t,\Delta), t \geq 0\}$ be a standard Brownian motion with drift $\Delta > 0$. Let*

$$\begin{aligned} l(t) &= -a + \lambda t \\ u(t) &= a - \lambda t \end{aligned}$$

for some $a > 0$ and $\lambda = \Delta/(2b)$ for some positive integer b. Let $c(t)$ denote the continuation region $(l(t), u(t))$ and let T be the first time that $\mathcal{B}(t,\Delta) \notin c(t)$. Then

$$\begin{aligned} \Pr\{\mathcal{B}(T,\Delta) < 0\} \quad \leq \quad & \sum_{j=1}^{b} (-1)^{j+1} \left(1 - \frac{1}{2}\mathcal{I}(j=b)\right) \\ & \times \exp\{-2a\lambda(2b-j)\,j\}. \end{aligned}$$

Fabian's bound on Pr{ICS} is particularly useful because a is the term that depends on the sample variance (see Paulson's a in Equation (8) for intuition). Thus, appropriately standardized, $\exp(-a)$ is related to the moment generating function of a chi-squared random variable, which simplifies unconditioning on the sample variance.

3.5 Large Number of Alternatives

The number of alternatives of interest in simulation problems can be quite large, with up to 100 being relatively common. However, Bechhofer-like IZ procedures were developed for relatively small numbers of alternatives, say no more than 20. They can be inefficient when the number of alternatives is large because they were developed to protect against the LFC—the configuration of system means under which it is most difficult to correctly select the best—to free the procedure from dependence on the true differences among the means. The Slippage Configuration (SC), $\mu_i = \mu_k - \delta$ for $i = 1, 2, \ldots, k-1$, is known to be the LFC for many procedures.

When the number of systems is large we rarely encounter anything remotely like the SC configuration, because large numbers of alternatives typically result from taking all feasible combinations of some controllable decision variables. Thus, the performance measures of the systems are likely to be spread out, rather than all clustered near the best. Paulson-like procedures with elimination might seem to be a cure for this ill, but the inequalities used to decompose the problem of k systems into paired comparisons with system k are typically quite conservative and become much more so with increasing k (although Kim and Nelson's (2001) fully sequential procedure $\mathcal{KN}$, described in the next section, has been shown to work well for up to $k = 500$ systems).

To overcome the inefficiency of IZ approaches for large numbers of alternatives, one idea is to try to gain the benefits of screening, as in Paulson-like procedures, but avoid the conservatism required to compensate for so many looks at the data. Nelson, et al. (2001) proposed spending some of the α for incorrect selection on an initial screening stage (using a Gupta-like subset-selection procedure), and spending the remainder on a second ranking stage (using a Bechhofer-like IZ procedure). Additive and multiplicative α spending is possible, depending on the situation (see Nelson, et al. 2001 and Wilson 2001). The resulting procedure, named NSGS, is presented in the next section.

This so-called "α-spending" approach—spreading the probability of incorrect selection across multiple stages—is a general-purpose tool, and there is no inherent reason to use only a single split. See Jennison and Turnbull (2000) for a thorough discussion.

4 EXAMPLE PROCEDURES

In this section we present two specific procedures to illustrate the concepts described in earlier sections. The NSGS procedure, due to Nelson, et al. (2001), and the $\mathcal{KN}$ procedure, due to Kim and Nelson (2001), are appropriate for terminating simulations or for steady-state simulations when multiple replications are employed.

The NSGS procedure requires that the output data from each system are i.i.d. normal, and that outputs across systems are independent, which leaves out CRN. NSGS is the combination of a Gupta-like subset-selection procedure, to reduce the number of alternatives still in play after the first stage of sampling, and a Bechhofer-like ranking procedure applied to the systems in the subset. The procedure uses α-spending between the subset selection and ranking to control the overall PCS. Banerjee's inequality allows the subset selection procedure to handle unequal variances.

4.1 Procedure NSGS

1. Specify the overall desired probability of correct selection $1-\alpha$, the IZ parameter δ, a common initial sample size from each system $n_0 \geq 2$, and the initial number of competing systems k. Further, set
$$t = t_{n_0-1,1-(1-\alpha/2)^{\frac{1}{k-1}}}$$
and obtain Rinott's constant $h = h(n_0, k, 1-\alpha/2)$ from the tables in Wilcox (1984) or Bechoffer et al. (1995). See also Table 8.3 in Goldsman and Nelson (1998).
2. Take n_0 outputs from each system. Calculate the first-stage sample means $\bar{X}_i(n_0)$ and marginal sample variances
$$S_i^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left(X_{ij} - \bar{X}_i(n_0)\right)^2,$$
for $i = 1, 2, \ldots, k$.

3. *Subset Selection.* Calculate the quantity

$$W_{i\ell} = t\left(\frac{S_i^2 + S_\ell^2}{n_0}\right)^{1/2}$$

for all $i \neq \ell$. Form the screening subset I, containing every alternative i such that $1 \leq i \leq k$ and

$$\bar{X}_i(n_0) \geq \bar{X}_\ell(n_0) - (W_{i\ell} - \delta)^+ \quad \text{for all } \ell \neq i.$$

4. If $|I| = 1$, then stop and return the system in I as the best. Otherwise, for all $i \in I$, compute the second-stage sample sizes

$$N_i = \max\left\{n_0, \lceil (hS_i/\delta)^2 \rceil\right\},$$

where $\lceil \cdot \rceil$ is the ceiling function.
5. Take $N_i - n_0$ additional outputs from all systems $i \in I$.
6. Compute the overall sample means $\bar{X}_i(N_i)$ for all $i \in I$. Select the system with the largest $\bar{X}_i(N_i)$ as best.

Nelson et al. (2001) showed that any subset-selection procedure and any two-stage IZ ranking procedure that satisfy certain mild conditions can be combined in this way while guaranteeing the overall probability of correct selection. The NGSG procedure can handle a relatively large number of systems because the first-stage screening is pretty tight. Nelson et al. (2001) provide a revised version of the NGSG procedure, the Group-Screening procedure, in which one can avoid simulating all the systems simultaneously. Boesel et al. (2003) extended the Group-Screening procedure for "clean up" after optimization via simulation.

The $\mathcal{KN}$ procedure is *fully sequential* because it takes only a single basic output from each alternative still in contention at each stage. Also, if there exists clear evidence that a system is inferior, then it will be eliminated from consideration immediately—unlike the NSGS procedure, where elimination occurs only after the first stage. $\mathcal{KN}$ also requires i.i.d. normal data, but does allow CRN. $\mathcal{KN}$ exploits the ideas of using paired differences, and controlling the Pr{ICS} on pairs to control it overall. Fabian's result is used to bound the error of a Brownian motion process that approximates each pair.

4.2 Procedure $\mathcal{KN}$

1. *Setup.* Select confidence level $1-\alpha$, IZ parameter δ and first stage sample size $n_0 \geq 2$. Set

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

2. *Initialization.* Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta(n_0-1)$. Obtain n_0 outputs X_{ij} $(j = 1, 2, \ldots, n_0)$ from each system i $(i = 1, 2, \ldots, k)$ and let $\bar{X}_i(n_0) = n_0^{-1}\sum_{j=1}^{n_0} X_{ij}$ denote the sample mean of the first n_0 outputs from system i.
For all $i \neq \ell$ compute

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

the sample variance of the difference between systems i and ℓ. Set $r = n_0$.
3. *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

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

4. *Stopping Rule.* If $|I| = 1$, then stop and select the system whose index is in I as the best.
Otherwise, take one additional output $X_{i,r+1}$ from each system $i \in I$, set $r = r + 1$ and go to *Screening*.

The $\mathcal{KN}$ procedure requires simulation of all systems simultaneously and a lot of switching among them. As discussed in Section 3, the switching cost can overwhelm the sampling cost, but this has become less of an issue in modern computing environments.

5 APPLICATION

This section illustrates NSGS and $\mathcal{KN}$ using an (s, S) inventory system with the five inventory policies as described in Koenig and Law (1985). The goal of this study is to compare the five polices given in Table 1 and find the one with the smallest expected average cost per month for the first 30 months of operation. Table 1 also contains the expected cost (in thousands of dollars) of each policy, which can be analytically computed in this case. We set $\delta = \$1$ thousand, $n_0 = 10$ initial replications, and $1 - \alpha = 0.95$.

Table 2 shows the results of the simulation study for each procedure, including the total number of outputs taken and the sample average cost per month for each policy. In NSGS, policies 3, 4, and 5 were eliminated after the first stage of sampling, so only policies 1 and 2 received second-stage samples. In $\mathcal{KN}$, only policies 4 and 5 were eliminated after the first stage, but the elimination of policies

Table 1: The Five Alternative Inventory Policies

Policy i	s	S	Expected Cost
1	20	40	114.176
2	20	80	112.742
3	40	60	130.550
4	40	100	130.699
5	60	100	147.382

Table 2: Simulation Results of the (s, S) Inventory Policy Example

	NSGS		$\mathcal{KN}$	
Policy i	# Obs.	Average	# Obs.	Average
1	209	114.243	98	114.274
2	349	112.761	98	113.612
3	10	130.257	16	130.331
4	10	128.990	10	128.990
5	10	147.133	10	147.133
Total	588		232	

3 and 1 occurred after they received 16 and 98 observations, respectively. This illustrates the value of the tighter initial screen in NSGS, which takes only one look at the data, and the potential savings from taking many looks, as $\mathcal{KN}$ does. Both procedures chose policy 2 as the best (which is in fact correct). Since δ is smaller than the true difference, NSGS and $\mathcal{KN}$ will choose the true best with 95% confidence. However, in general we do not have any information about the true differences; therefore, the best we can conclude without prior knowledge is that policy 2 is either the true best, or has expected cost per month within \$1 thousand of the true best policy, with 95% confidence.

6 ASYMPTOTIC VALIDITY

When normality and independence of the output from within each system are untenable assumptions, proving that R&S procedures provide a correct-selection guarantee for a finite sample is largely hopeless. Nevertheless, well designed procedures have shown good empirical performance. *Asymptotic analysis* can provide theoretical support for this observation. Asymptotic analysis typically means analysis as the simulation effort (run length, number of replications, or perhaps both) increases (conceptually) without bound. The power of asymptotic analysis is that many of the problem-specific details that thwart mathematical analysis in the finite-sample case wash out in the limit. Asymptotic analysis, done appropriately, can establish conditions under which we can expect procedures to work, rather than just relying on limited empirical evidence that they do; it can also establish the asymptotic superiority of one procedure over another.

We mention two useful regimes for asymptotic analysis of R&S procedures:

- Dealing with non-normal or dependent data: In this regime, the goal is to show that a selection procedure does guarantee the PCS requirement if enough simulation effort is expended. One example is Kim and Nelson (2003), who drive the run length to infinity by letting both the indifference zone δ and the true differences among the systems' means go to 0, so that the PCS approaches a meaningful limit, rather than 1. We can interpret their result as telling us what will happen as the problem becomes more and more difficult, which is what we would like to know since few errors occur in easy problems where the means are dramatically different.
- Comparing procedures: The variance of any sensible point estimator will go to zero as the sample size goes to infinity, but that does not mean that all point estimators are equally good. Scaling up the variance at the same rate at which it is going to zero can sometimes reveal important differences among estimators. Similarly, we can look at the rate at which the simulation effort of an IZ procedure increases as $P^* \to 1$ and compare the rates of competing procedures to establish asymptotic superiority of one over another. See, for instance, Jennison, Johnstone and Turnbull (1982).

7 OTHER FORMULATIONS

Throughout this paper we have focused on the problem of finding the best when the best is defined as the system with the largest or smallest expected performance measure. As discussed in Section 1, there exist other types of comparison problems. Here we briefly visit each type of comparison problem and provide useful references.

1. *Comparisons with a standard*: The goal of comparison with a standard is to find systems whose expected performance measures are larger (smaller) than a standard and, if there are any, to find the one with the largest (smallest) expected performance. For this type of problem, each alternative needs to be compared to the standard as well as other alternative systems. Nelson and Goldsman (2001) proposed two-stage procedures and Kim (2002) proposed fully sequential procedures.
2. *Multinomial selection*: In multinomial selection problem, the definition of the best is the system that is *mostly likely* to be the best. In the simulation context this typically means identifying the system i with the largest value of p_i, where

$p_i = \Pr\{X_{ij} > X_{\ell j}, \forall \ell \neq i\}$ for a maximization problem. Bechhofer, Elmaghraby, and Morse (BEM) (1959) proposed a single-stage procedure that finds the most likely system while meeting a certain IZ criterion. With simulation in mind, Miller et al. (1998) devised another single-stage procedure that achieves a higher probability of correct selection than does BEM.

3. *Bernoulli selection*: In Bernoulli selection problems, the basic output from each system on each replication is either one ("success") or zero ("fail") and the best is defined as the system with the largest probability of success. See Chapter 7 of Bechhofer, et al. (1995) for a comprehensive reference.
4. *Bayesian approach:* Instead of providing a PCS guarantee, Bayesian approaches attempt to allocate a finite data budget to maximize the posterior PCS of the selected system. Chen, et al. (2000) and Chick and Inoue (2001) are two recent references.

REFERENCES

Banerjee, S. 1961. On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t-table. *Sankhyā* A23:359–378.

Banks, J., J. S. Carson, B. L. Nelson, and D. Nicol. 2001. *Discrete-Event System Simulation*. Upper Saddle River, NJ: Prentice Hall.

Bechhofer, R. E. 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* 25:16–39.

Bechhofer, R. E., S. Elmaghraby, and N. Morse. 1959. A single-sample multiple-decision procedure for selecting the multinomial event which has the highest probability. *Annals of Mathematical Statistics* 30:102–119.

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

Boesel, J., B. L. Nelson, and S.-H. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. To appear in *Operations Research*.

Chen, H. C., C. H. Chen, and E. Yücesan. 2000. Computing efforts allocation for ordinal optimization and discrete event simulation. *IEEE Transactions on Automatic Control* 45:960–964

Chick, S., and K. Inoue. 2001. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49:1609–1624.

Dudewicz, E. J. 1995. The heteroscedastic method: Fifty+ years of progress 1945–2000, and professor Minoru Siotani's award-winning contributions. *American Journal of Mathematical and Management Sciences* 15:179–197.

Fabian, V. 1974. Note on Anderson's sequential procedures with triangular boundary. *Annals of Statistics* 2:170–176.

Glynn, P. W., and D. L. Iglehart. 1990. Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15:1–16.

Goldsman, D., and B. L. Nelson. 1998. Comparing systems via simulation. In *Handbook of Simulation*, ed. J. Banks, 273–306. New York: John Wiley.

Gupta, S. S. 1956. On a decision rule for a problem in ranking means. Doctoral dissertation, Institute of Statistics, Univ. of North Carolina, Chapel Hill, NC.

Gupta, S. S. 1965. On some multiple decision (ranking and selection) rules. *Technometrics* 7:225–245.

Hartmann, M. 1988. An improvement on Paulson's sequential ranking procedure. *Sequential 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. *Sequential Analysis* 10:1–16.

Hochberg, Y., and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. New York: John Wiley.

Hong, L. J., and B. L. Nelson. 2003. The tradeoff between sampling and switching: New sequential procedures for indifference-zone selection. Technical Report, Dept. of IEMS, Northwestern Univ., Evanston, IL.

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

Jennison, C., I. M. Johnstone, and B. W. Turnbull. 1980. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. Technical Report, Dept. of ORIE, Cornell Univ., Ithaca, NY.

Jennison, C., I. M. Johnstone, and B. W. Turnbull. 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*, ed. S. S. Gupta and J. O. Berger, 55–86. New York: Academic Press.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman & Hall.

Koenig, L. W., and A. M. Law. 1985. A procedure for selecting a subset of size m containing the ℓ best of k independent normal populations, with applications to simulation. *Communications in Statistics—Simulation and Computation* B14:719–734.

Kim, S.-H. 2002. Comparison with a standard via fully sequential procedures. Technical Report, School of ISyE, Georgia Tech, Atlanta, GA.

Kim, S.-H., and B. L. Nelson. 2001. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS* 11:251–273.

Kim, S.-H., and B. L. Nelson. 2003. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. Technical Report, Dept. of IEMS, Northwestern Univ., Evanston, IL.

Kimball, A. W. 1951. On dependent tests of significance in the analysis of variance. *Annals of Mathematical Statistics* 22:600–602.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*, 3d ed. New York: McGraw-Hill.

Miller, J. O., B. L. Nelson, and C. H. Reilly. 1998. Efficient multinomial selection in simulation. *Naval Research Logistics* 45:459–482.

Nelson, B. L., and D. Goldsman. 2001. Comparisons with a standard in simulation experiments, *Management Science* 47:449–463.

Nelson, B. L., and F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* 41:1935–1945

Nelson, B. L., J. Swann, D. Goldsman, and W.-M. T. Song. 2001. Simple procedures for selecting the best system when the number of alternatives is large. *Operations Research* 49:950–963.

Paulson, E. 1964. A sequential procedure for selecting the population with the largest mean from *k* normal populations. *Annals of Mathematical Statistics* 35:174–180.

Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics—Theory and Methods* A7:799–811.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag.

Slepian, D. 1962. The one-sided barrier problem for Gaussian noise. *Bell Systems Technical Journal* 41:463–501.

Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 16:243–258.

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

Wald, A. 1947. *Sequential Analysis.* New York: John Wiley.

Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 25:350–362.

Wilcox, R. R. 1984. A table for Rinott's selection procedure. *Journal of Quality Technology* 16:97–100.

Wilson, J. R. 2001. A multiplicative decomposition property of the screening-and-selection procedures of Nelson et al. *Operations Research* 49:964–966.

AUTHOR BIOGRAPHIES

SEONG-HEE KIM is an Assistant Professor in the School of Industrial Systems and Engineering at Georgia Tech. Her research interests include simulation output analysis and ranking and selection. Her e-mail and web addresses are <skim@isye.gatech.edu> and <www.isye.gatech.edu/~skim/>.

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