SIMPLE PROCEDURES FOR SELECTING THE BEST SIMULATED SYSTEM WHEN THE NUMBER OF ALTERNATIVES IS LARGE

BARRY L. NELSON

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208-3119, nelsonb@northwestern.edu

JULIE SWANN

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, julie.swann@isye.gatech.edu

DAVID GOLDSMAN

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, sman@isye.gatech.edu

WHEYMING SONG

Department of Industrial Engineering, National Tsing Hua University, Hsinchu R.O.C., Taiwan, wheyming@ie.nthu.edu.tw

(Received January 1998; revisions received January 1999, November 1999; accepted July 2000)

In this paper, we address the problem of finding the simulated system with the best (maximum or minimum) expected performance when the number of alternatives is finite, but large enough that ranking-and-selection (R&S) procedures may require too much computation to be practical. Our approach is to use the data provided by the first stage of sampling in an R&S procedure to screen out alternatives that are not competitive, and thereby avoid the (typically much larger) second-stage sample for these systems. Our procedures represent a compromise between standard R&S procedures—which are easy to implement, but can be computationally inefficient—and fully sequential procedures—which can be statistically efficient, but are more difficult to implement and depend on more restrictive assumptions. We present a general theory for constructing combined screening and indifference-zone selection procedures, several specific procedures and a portion of an extensive empirical evaluation.

1. INTRODUCTION

A central reason for undertaking many—perhaps most—stochastic simulation studies is to find a system design that is the best, or near the best, with respect to some measure or measures of system performance. The statistical procedure that is most appropriate for this purpose depends on the characteristics of the problem at hand, characteristics that include the number of alternative designs, the number of performance measures, whether or not the alternatives are functionally related in some useful way, and what, if any, regularity conditions apply to the response surface. Comprehensive reviews of the available tools can be found in Fu (1994) and Jacobson and Schruben (1989).

When the number of alternative designs is relatively small, say 2 to 10, and there is not a strong functional relationship among them, then statistical procedures based on the theory of ranking and selection (R&S) are popular because they are easy to apply and interpret. See, for instance, Bechhofer et al. (1995) for a treatment of the general topic of R&S, and Goldsman and Nelson (1998) for a survey of R&S procedures applied to simulation. When mean performance is of interest, the typical *indifference-zone* (IZ) selection procedure conforms to the following recipe:

1. For each alternative, obtain a (usually small) number of observations of the system performance measure of interest and calculate a measure of the variability of the observations.

2. Based on the measure of variability, the number of alternatives, and the desired confidence level, determine the total number of observations needed from each alternative to guarantee that a user-specified practically significant difference in performance can be detected at the desired confidence level.

3. Obtain the prescribed number of additional observations from each alternative and select the one with the best sample performance.

Why are IZ selection procedures well suited for the simulation environment? First of all, many of these procedures assume that the data from a particular competitor are independent and normally distributed—assumptions that are roughly satisfied by appropriately batched data or by sample averages of independent replications of the simulations.

Subject classifications: Simulation, design of experiments: two-stage procedures. Simulation, statistical analysis: finding the best alternative. Statistics, design of experiments.
Area of review: SIMULATION.

Second, we can manipulate the underlying pseudorandom number seeds to produce simulations that are also independent *between* competitors, possibly running the different competitors on parallel processors.

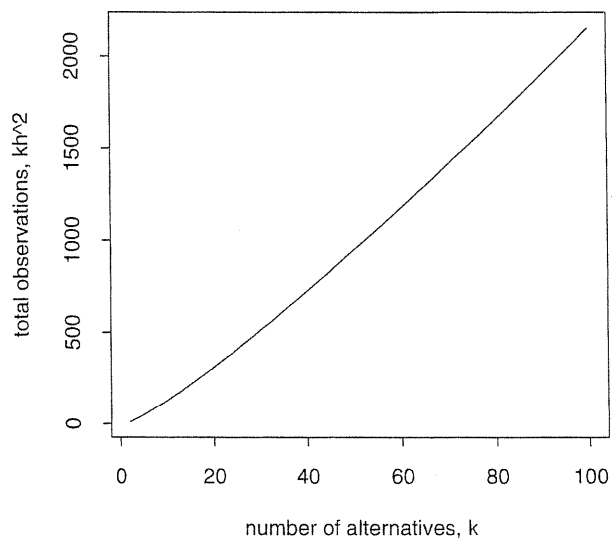
On the other hand, why are such procedures only recommended for a small number of alternatives? Consider Step 2 for a specific procedure due to Rinott (1978). Rinott's procedure specifies N_i, the total number of independent, normally distributed observations required from alternative i, to be

$$N_i = \max\left\{n_0, \left\lceil \left(\frac{hS_i}{\delta}\right)^2 \right\rceil\right\}, \quad (1)$$

where n_0 is the initial sample size; h is a constant that depends on the number of alternatives k, the desired confidence level $1-\alpha$, and n_0; S_i^2 is the sample variance of the initial n_0 observations; and δ is the practically significant difference specified by the user. Two features of Equation (1) argue against using it when the number of alternatives is large: (a) the constant h is an *increasing* function of k, and (b) the formula is based on the worst-case assumption that the true mean of the best alternative is exactly δ better than all of the others, and all of the others are tied for second best. This assumption is "worst case" in the sense that it makes the best alternative as hard as possible to separate from the others, given that it is at least δ better than anything else. The reason for assuming the worst case is that it allows formula (1) to be independent of the true or sample means.

The focus of this paper is issue (b). However, it is worth noting that the growth of h as a function of k is typically quite slow. Figure 1 plots kh^2 versus k for Rinott's procedure when the confidence level for correctly selecting the best alternative is 95%, and the first-stage sample size is large. Because N_i, the number of observations required from alternative i, is proportional to h^2, the figure shows the expected total number of observations in the experiment in units of $(\sigma/\delta)^2$ (under the assumption that the variances across all alternatives equal σ^2). Notice that within the range of $k = 2$ to 100 systems, each additional system adds approximately $20(\sigma/\delta)^2$ observations to the experiment; in other words, the computational effort increases linearly in the number of systems. Thus, issue (a) is not as serious as it is often thought to be.

Figure 1. The expected total sample size kh^2 in units $(\sigma/\delta)^2$ for Rinott's procedure as a function of the number of systems k for 95% confidence of making a correct selection.



The fact that IZ selection procedures are based on a worst-case analysis implies that they may prescribe more observations than needed in order to deliver the desired correct-selection guarantees.[1] This is especially unfortunate when the number of alternatives is large and they differ widely in performance, because a great deal of simulation effort may be wasted on alternatives that are not competitive with the best. One situation in which a large number of heterogeneous alternatives may arise is after termination of a stochastic optimization procedure. Such procedures either provide no statistical guarantee or only a guarantee of asymptotic convergence. We have proposed using R&S to "clean up" after stochastic optimization—ideally with minimal additional sampling—to insure that the alternative selected as best is indeed the best or near the best among all of those visited by the search (Boesel and Nelson 1998, Boesel et al. 2001).

Our goal is to provide procedures that are more adaptive than standard IZ selection procedures without losing the ease of implementation and interpretation that make them attractive. Specifically, we will provide simple screening procedures that can be used to eliminate noncompetitive systems after Step 1, thereby saving the (possibly large) number of observations that would be taken at Step 3. The screening procedures we propose are based on the *subset selection* branch of R&S. These procedures attempt to select a (possibly random-size) subset of the k competing systems that contains the one with the largest or smallest expected performance. Gupta (1956, 1965) proposed a single-stage procedure for this problem that is applicable when the samples from the competing alternatives are independent, equal-sized, and normally distributed with common unknown variance. The fact that subset selection can be done in a single stage of sampling is a feature we exploit, but we first need to extend the existing methods to allow for unknown and unequal variances across systems. Unknown and unequal variance subset-selection procedures do exist, but they require two or more stages of sampling in and of themselves (see, for instance, Koenig and Law 1985 and Sullivan and Wilson 1989).

In this paper we combine subset selection—to screen out noncompetitive systems—with IZ selection—to select the best from among the survivors of screening—to obtain a computationally and statistically efficient procedure. Procedures of this type have appeared in the literature before, e.g., Gupta and Kim (1984), Hochberg and Marcus (1981), Santner and Behaxeteguy (1992), Tamhane (1976, 1980),

and Tamhane and Bechhofer (1977, 1979). One purpose of our paper is to extend this work, which typically assumes known variances, or unknown but equal variances, so as to be more useful in simulation experiments.

If the idea of sample-screen-sample-select is effective, then one might naturally push it to the limit to obtain a *fully sequential procedure* (see, for instance, Hartmann 1991 and Paulson 1964). Such procedures take a single observation from each alternative that is still in play at the current stage of sampling, eliminate noncompetitive systems, then continue with a single observation from each remaining alternative, and so on.

One disadvantage of a fully sequential procedure is the overhead required to repeatedly switch among alternative simulated systems to obtain a vector of observations across all systems still in play. Another is that existing procedures require equal variances across systems, an assumption that is clearly violated in many systems-simulation problems (in related research we are working on removing this restriction). Finally, the situations in which fully sequential procedures such as Hartmann (1991) and Paulson (1964) tend to beat screen-and-select procedures occur when all of the alternatives are close in performance (Bechhofer et al. 1990), while we are interested in situations characterized by a large number of alternatives with heterogeneous performance. On the whole, though, we believe that the use of sequential procedures is ultimately a good idea—the only problems lie in overcoming the above obstacles. Thus, we are also pursuing the development of such procedures (Kim and Nelson 2001).

The paper is organized as follows: §2 presents a decomposition lemma that allows us to marry screening procedures to IZ selection procedures while maintaining overall statistical error control. In §3 we extend the state of the art in single-stage screening to allow unequal variances across alternatives; then in §4, we combine these new screening procedures with IZ-selection procedures for the case of unequal variances. Section 5 extends our ideas to allow the systems to be screened in groups, rather than all at once, which is convenient in an exploratory study or in conjunction with an optimization/search algorithm, and can be more efficient. The paper ends with a report on a large-scale empirical study in §6, and conclusions in §7.

2. A DECOMPOSITION LEMMA

In this section we present a key lemma that simplifies the construction of combined screening and IZ selection procedures. The Bonferroni-like lemma, which generalizes a result in Hochberg and Marcus (1981), establishes that under very general conditions we can apply an IZ selection procedure to the survivors of a screening procedure and still guarantee an overall probability of correct selection (CS) even if the selection procedure starts with the same data that was used for screening.

Let the alternative systems be numbered $1, 2, \ldots, k$ and let $[k]$ denote the unknown index of the best alternative. Suppose $\mathcal{S}$ is a procedure that obtains data from each system, and based on that data determines a (possibly) random subset I of $\{1, 2, \ldots, k\}$, such that $\Pr\{\mathcal{A}\} \geq 1 - \alpha_0$, where $\mathcal{A} = \{[k] \in I\}$ is the event that I contains the best alternative.

We want to consider combining $\mathcal{S}$ with a multiple-stage selection procedure $\mathcal{R}$ that will take I—and the data used to determine I—as its initial stage of sampling, and then obtain additional data in an attempt to determine the best system. The types of selection procedures we have in mind can be applied independently of any screening procedure or, equivalently, can be viewed as retaining all systems after the first stage of sampling.

Let $J_\ell, \ell = 1, 2, \ldots, s$ be the distinct subsets of $\{1, 2, \ldots, k\}$ that contain $[k]$. There are $s = 2^{k-1}$ such subsets. We require that $\mathcal{R}$ be a procedure that has the following properties: $\mathcal{R}$ determines a (possibly random) index $K \in J_\ell$ such that $\Pr\{\mathcal{B}(J_\ell)\} \geq 1 - \alpha_1$ for *any* such subset J_ℓ that contains $[k]$, where $\mathcal{B}(J_\ell) = \{K = [k]\}$ is the event that K is the best alternative. Further, suppose that $\mathcal{B}(\{1, 2, \ldots, k\}) \subseteq \mathcal{B}(J_\ell)$ for all J_ℓ. In other words, if a correct selection would be made when $\mathcal{R}$ is applied to the entire set, then a correct selection would be made if it were applied to any subset that contains $[k]$. This property will hold for any procedure whose sampling from system i, say, depends only on the data generated for system i. This will be true, for instance, when $\mathcal{R}$ takes an initial sample of n_0 observations from system i, then determines the total sample from system i by a formula like

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

where S_i^2 is the sample variance of the initial n_0 observations from system i and both δ or h are fixed.

Let $\mathcal{B} = \bigcap_{\ell=1}^{s} \mathcal{B}(J_\ell)$, the event that $[k]$ is selected for all subsets J_ℓ to which $\mathcal{R}$ can be applied.

LEMMA 1. *For the combined procedure* $\mathcal{S} + \mathcal{R}$,

$$\Pr\{\text{CS}\} \geq \Pr\{\mathcal{A} \cap \mathcal{B}\} \geq 1 - (\alpha_0 + \alpha_1).$$

PROOF. Any outcome belonging to the event $\mathcal{B}$ results in a correct selection, provided that the subset of systems considered by $\mathcal{R}$ contains $[k]$. The event $\mathcal{A}$ only provides outcomes for which this is the case. Any outcome that satisfies both conditions will certainly result in a correct selection.

Next notice that

$$\begin{aligned}\Pr\{\mathcal{A} \cap \mathcal{B}\} &= \Pr\{\mathcal{A}\} + \Pr\{\mathcal{B}\} - \Pr\{\mathcal{A} \cup \mathcal{B}\} \\ &\geq \Pr\{\mathcal{A}\} + \Pr\{\mathcal{B}(\{1, \ldots, k\})\} - \Pr\{\mathcal{A} \cup \mathcal{B}\} \\ &\geq (1 - \alpha_0) + (1 - \alpha_1) - 1,\end{aligned}$$

where the first inequality follows because $\Pr\{\bigcap_{\ell=1}^{s} \mathcal{B} \times (J_\ell)\} \geq \Pr\{\mathcal{B}(\{1, 2, \ldots, k\})\}$. □

REMARK. The additive decomposition of the screening and IZ selection procedures in Lemma 1 will not be the most statistically efficient in all cases. For some specific procedures, we have shown that $\Pr\{\mathcal{A} \cap \mathcal{B}\} \geq (1 - \alpha_0)(1 - \alpha_1)$.

Notice that $1-(\alpha_0+\alpha_1)-(1-\alpha_0)(1-\alpha_1)=-\alpha_0\alpha_1<0$, implying that the additive decomposition is more conservative (and therefore less statistically efficient) than the multiplicative one. However, the difference is quite small—in the third decimal place or beyond for standard confidence levels.

3. SCREENING PROCEDURES

The decomposition lemma allows us to derive screening procedures in isolation from the selection procedures with which they will be combined. In this section we present a new screening procedure that yields a subset of random size that is guaranteed to contain $[k]$ with probability $\geq 1-\alpha_0$. This procedure generalizes others in the literature by permitting unequal variances, which certainly is the case in many simulation studies. The procedure also exploits the dependence induced by the use of common random numbers.

We will use the following notation throughout the paper: Let X_{ij} be the jth observation from alternative i, for $i=1,2,\ldots,k$. We will assume that the X_{ij} are i.i.d. $\mathrm{N}(\mu_i,\sigma_i^2)$ random variables, with both μ_i and σ_i^2 unknown and (perhaps) unequal. These assumptions are reasonable when the output data are themselves averages of a large number of more basic output data, either from different replications or from batch means within a single replication. The ordered means are denoted by $\mu_{[1]}\leq\mu_{[2]}\leq\cdots\leq\mu_{[k]}$. We assume that bigger is better, implying that the goal of the experiment is to find the system associated with $\mu_{[k]}$. In other words, we define a CS to mean that we select a subset that contains the index $[k]$. We will require that the $\Pr\{\mathrm{CS}|\mu_{[k]}-\mu_{[k-1]}\geq\delta\}\geq 1-\alpha$, where "$\mu_{[k]}-\mu_{[k-1]}\geq\delta$" is a reminder that the guarantee is valid, provided the difference between the best and second-best true mean is at least δ. Throughout the paper, we use $[i]$ to denote the unknown index of the system with the ith smallest mean.

The following procedure can be applied when an initial sample having common size has been obtained from all systems. No further sampling is required, but as a result the size of the subset is random.

Screen-to-the-Best Procedure.

1. Select the overall confidence level $1-\alpha_0$, practically significant difference δ, sample size $n_0\geq 2$, and number of systems k. Set $t=t_{(1-\alpha_0)^{\frac{1}{k-1}},n_0-1}$, where $t_{\beta,\nu}$ denotes the β quantile of the t distribution with ν degrees of freedom.
2. Sample X_{ij}, $i=1,2,\ldots,k$; $j=1,2,\ldots,n_0$.
3. Compute the sample means and variances $\overline{X}_i$ and S_i^2 for $i=1,2,\ldots,k$. Let

$$W_{ij}=t\left(\frac{S_i^2}{n_0}+\frac{S_j^2}{n_0}\right)^{1/2}$$

for all $i\neq j$.

4. Set $I=\{i: 1\leq i\leq k \text{ and } \overline{X}_i\geq\overline{X}_j-(W_{ij}-\delta)^+, \forall\, j\neq i\}$, where $y^+=\max\{0,y\}$.
5. Return I.

The subset I can be thought of as containing those alternatives whose sample means are not significantly inferior to the best of the rest. In addition, notice that the subset I will never be empty for this procedure. In the appendix we prove that $\Pr\{[k]\in I \mid \mu_{[k]}-\mu_{[k-1]}\geq\delta\}\geq 1-\alpha_0$. In the special case of $\delta=0$ this is a generalization of Gupta's (1965) subset-selection procedure to allow unequal variances.

REMARK. If common random numbers (CRN) are used to induce dependence across systems, then the subset-selection procedure described above is still valid, provided $S_i^2/n_0+S_j^2/n_0$ is replaced by S_{ij}^2/n_0, where

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

and $t=t_{1-\alpha_0/(k-1),n_0-1}$.

4. COMBINED PROCEDURES

The decomposition lemma makes it is easy to apply an IZ selection procedure to the systems retained by a screening procedure, while still controlling the overall confidence level. The key observations are as follows:

- For overall confidence level $1-\alpha$, choose confidence levels $1-\alpha_0$ for the screening procedure, and $1-\alpha_1$ for the IZ selection procedure such that $\alpha_0+\alpha_1=\alpha$. A convenient choice is $\alpha_0=\alpha_1=\alpha/2$.
- Choose the critical constant t for the screening procedure to be appropriate for k systems, n_0 initial observations, and confidence level $1-\alpha_0$.
- Choose the critical constant h for the IZ selection procedure to be appropriate for k systems, n_0 initial observations, and confidence level $1-\alpha_1$.

Below we exhibit one such procedure that combines the screening procedure of §3 with Rinott's IZ selection procedure.

Combined Procedure.

1. Select overall confidence level $1-\alpha$, practically significant difference δ, first-stage sample size $n_0\geq 2$, and number of systems k. Set $t=t_{(1-\alpha_0)^{\frac{1}{k-1}},n_0-1}$ and $h=h(1-\alpha_1,n_0,k)$, where h is Rinott's constant (see Wilcox 1984 or Bechhofer et al. 1995 for tables), and $\alpha_0+\alpha_1=\alpha$.
2. Sample X_{ij}, $i=1,2,\ldots,k$; $j=1,2,\ldots,n_0$.
3. Compute the first-stage sample means and variances $\overline{X}_i^{(1)}$ and S_i^2 for $i=1,2,\ldots,k$. Let

$$W_{ij}=t\left(\frac{S_i^2}{n_0}+\frac{S_j^2}{n_0}\right)^{1/2}$$

for all $i\neq j$.

4. Set $I=\{i: 1\leq i\leq k \text{ and } \overline{X}_i^{(1)}\geq\overline{X}_j^{(1)}-(W_{ij}-\delta)^+, \forall\, j\neq i\}$.
5. If I contains a single index, then stop and return that system as the best.

Otherwise, for all $i\in I$ compute the second-stage sample size

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

6. Take $N_i - n_0$ additional observations from all systems $i \in I$ and compute the overall sample means

$$\overline{X}_i^{(2)} = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}$$

for $i \in I$.

7. Select as best the system $i \in I$ with the largest $\overline{X}_i^{(2)}$.

The conclusion that $\Pr\{\text{CS}|\mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq 1 - \alpha$ for the combined procedure follows immediately from Lemma 1 because the screening procedure is calibrated to retain the best with probability $\geq 1 - \alpha_0$, while Rinott's selection procedure guarantees a probability of correct selection $\geq 1 - \alpha_1$ for k (or fewer) systems. Thus, the overall probability of correct selection is $\geq 1 - (\alpha_0 + \alpha_1) = 1 - \alpha$.

REMARK. The goal of a procedure like the one above is to find the best system whenever it is at least δ better than any other. However, we also obtain statistical inference that holds no matter what the configuration of the true means. We claim that with probability greater than or equal to $1 - \alpha$ all of the following hold simultaneously:

- For all $i \in I^c$, we have $\mu_i < \max_{j \in I} \mu_j + \delta$; that is, we can claim with high confidence that systems excluded by screening are less than δ better than the best in the retained set. If we use $\delta = 0$ for screening, this means that we can claim with high confidence that the systems excluded by screening are not the best.
- For all $i \in I$,

$$\mu_i - \max_{j \in I,\, j \neq i} \mu_j \in \left[-\left(\overline{X}_i^{(2)} - \max_{j \in I,\, j \neq i} \overline{X}_j^{(2)} - \delta \right)^-, \left(\overline{X}_i^{(2)} - \max_{j \in I,\, j \neq i} \overline{X}_j^{(2)} + \delta \right)^+ \right],$$

where $-y^- = \min\{0, y\}$ and $y^+ = \max\{0, y\}$. These confidence intervals bound the difference between each alternative and the best of the others in I.

- If K denotes the index of the system that we select as the best, then

$$\Pr\left\{ \mu_K - \max_{i \in I,\, i \neq K} \mu_i \geq -\delta \right\} \geq 1 - \alpha.$$

In words, the system we select will be within δ of the best system in I with high confidence.

The first claim follows because the screening procedure is in fact one-sided multiple comparisons with a control with each system i taking a turn as the control (Hochberg and Tamhane 1987). The second claim follows from Proposition 1 of Nelson and Matejcik (1995), while the third claim follows from Nelson and Goldsman (2001), Corollary 1.

REMARK. Because we prefer working with sample means as estimators of the true system means, the procedure presented here is based on sample means. However, many subset-selection and IZ selection procedures have been based on *weighted* sample means, and these procedures are typically more efficient than corresponding procedures based on sample means. Examples include the restricted subset-selection procedures (RSSPs)[2] of Koenig and Law (1985), Santner (1975), and Sullivan and Wilson (1989); and the IZ selection procedure of Dudewicz and Dalal (1975). The decomposition lemma in §2 allows us to form combined procedures based on weighted sample means if we desire. Two obvious combinations are:

- Combine the screening procedure of §3 with the IZ selection procedure of Dudewicz and Dalal (1975) to obtain a two-stage procedure for selecting the best.
- Combine either the RSSP of Koenig and Law (1985) or Sullivan and Wilson (1989) with the IZ selection procedure of Dudewicz and Dalal (1975) to obtain a three-stage procedure for selecting the best that bounds the number of systems that survive screening.

We investigate one of these combinations later in the paper.

5. A GROUP-SCREENING VERSION OF THE COMBINED PROCEDURE

One difficulty with the combined screening and IZ selection procedure presented in the previous section is that all k of the systems must receive first-stage sampling before proceeding to the second stage. In this section, we show that it is possible—and sometimes advantageous—to break the overall screening-and-selection problem into a screening-and-selection problem over smaller groups of systems. As we will show, if one or more very good alternatives are found early in the process, then they can be used to eliminate *a greater number* of noncompetitive systems than would be eliminated if all were screened at once.

To set up the procedure, let $G_1, G_2, \ldots, G_m$ be groups of systems such that $G_1 \cup G_2 \cup \cdots \cup G_m = \{1, 2, \ldots, k\}$, $G_i \cap G_j = \varnothing$ for $i \neq j$, $|G_i| \geq 1$ for all i and $|G_1| \geq 2$. When we screen the ℓth group in the experiment, the systems in G_ℓ will be screened with respect to each other and all systems retained from the previous groups. However, the systems retained from previous screenings have already received second-stage sampling, so screening in group ℓ is based on

$$W_{ij} = t\left(\frac{S_i^2}{\widetilde{N}_i} + \frac{S_j^2}{\widetilde{N}_j} \right)^{1/2},$$

where

$$\widetilde{N}_i = \begin{cases} n_0, & \text{if system } i \text{ has only received first-stage sampling,} \\ N_i, & \text{if system } i \text{ has received second-stage sampling.} \end{cases}$$

The potential savings occur because typically $N_i \gg n_0$, which shortens W_{ij}, providing a tighter screening procedure.

In the following description of the group-screening procedure, we let $\overline{X}_i^{(s)}$ denote the sample mean of all observations taken from system i through $s = 1$ or 2 stages of

sampling; and we use I_ℓ to denote the set of all systems that have survived screening after ℓ groups have been screened.

Group-Screening Procedure.

1. As in Step 1 of the combined procedure in §4.
2. Let $I_0 = \varnothing$.
3. Do the following for $\ell = 1, 2, \ldots, m$:

(a) Sample X_{ij}, $j = 1, 2, \ldots, n_0$, for all $i \in G_\ell$; compute $\overline{X}_i^{(1)}$ and S_i^2; and set $\widetilde{N}_i = n_0$.

Comment: G_ℓ is the current group of systems to be screened.

(b) Let $I_\ell = I^{\text{new}} \cup I^{\text{old}}$, where

$$I^{\text{new}} = \{i : i \in G_\ell, \overline{X}_i^{(1)} \geq \overline{X}_j^{(1)} - (W_{ij} - \delta)^+, \ \forall\, j \in G_\ell \text{ and } \overline{X}_i^{(1)} \geq \overline{X}_j^{(2)} - (W_{ij} - \delta)^+, \ \forall\, j \in I_{\ell-1}\},$$

and

$$I^{\text{old}} = \{i : i \in I_{\ell-1}, \overline{X}_i^{(2)} \geq \overline{X}_j^{(1)} - (W_{ij} - \delta)^+, \ \forall\, j \in G_\ell \text{ and } \overline{X}_i^{(2)} \geq \overline{X}_j^{(2)} - (W_{ij} - \delta)^+, \ \forall\, j \in I_{\ell-1}\}.$$

Comment: I^{new} is the set of newly screened systems that make it into the next screening, while I^{old} is the set of previously retained systems that survive another round. At the cost of some additional data storage I^{old} can simply be set to $I_{\ell-1}$ so that all systems that survive one round of screening survive to the end.

(c) For all $i \in I^{\text{new}}$ compute the second-stage sample size N_i based on Rinott's procedure, sample $N_i - n_0$ additional observations, and compute the second-stage sample mean $\overline{X}_i^{(2)}$. Set $\widetilde{N}_i = N_i$.

4. Select as best the system $i \in I_m$ with the largest sample mean $\overline{X}_i^{(2)}$.

In the appendix we show that $\Pr\{\text{CS} | \mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq 1 - 2\alpha_0 - \alpha_1$. However, we *conjecture* that $\Pr\{\text{CS} | \mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq 1 - \alpha_0 - \alpha_1 = 1 - \alpha$. In the online companion to this paper we show why we believe that the conjecture is true, *and henceforth, we operate under this assumption.*

REMARK. There are several contexts in which this type of procedure might be useful.

- Consider exploratory studies in which not all system designs of interest are initially known or available. Suppose the user can bound k so that the critical values can be computed. Then the study can proceed in a relatively informal manner, with new alternatives added to the study as they occur to the analyst, or as suggested by the performance of other alternatives. The analyst can terminate the study at any point at which acceptable performance has been achieved and still have the desired correct-selection guarantees.
- Certain heuristic search procedures, such as genetic algorithms, work with groups or "populations" of systems at each iteration. The group-screening procedure allows new alternatives generated by the search to be compared with the best alternatives previously visited, while maintaining the overall confidence level for the system finally chosen (provided, again, that the user can bound k in advance). See, for instance, Boesel et al. (2001).

EXAMPLE. To illustrate the potential savings from group screening, consider a situation with $k = 3$ systems having the following characteristics:

	System		
	1	2	3
μ	3	0	1
σ^2	10	10	10

If the first-stage sample size is $n_0 = 10$, and we want an overall confidence level of $1 - \alpha = 0.95$, then the required critical values are $t_{(0.975)^{1/2}, 9} = 2.68$ and $h(0.975, 10, 3) = 3.72$.

Suppose that the practically significant difference is $\delta = 1$, and (for simplicity) the sample averages and sample variances perfectly estimate their population counterparts. Then, when all systems are screened at once we have

$$W_{ij} = 2.68\left(\frac{10}{10} + \frac{10}{10}\right)^{1/2} \approx 3.79 \quad \text{for all } i \neq j.$$

Clearly, System 1 survives because it is the sample best, System 2 is screened out, but System 3 also survives because

$$\overline{X}_3^{(1)} = 1 > \overline{X}_1^{(1)} - (W_{31} - \delta)^+ = 3 - (3.79 - 1)^+ = 0.21.$$

Thus, both Systems 1 and 3 receive second-stage sampling, in which case

$$N_i = \left\lceil \left(\frac{3.72\sqrt{10}}{1}\right)^2 \right\rceil = 139 \quad \text{for } i = 1, 3.$$

Now suppose that the systems are grouped as $G_1 = \{1, 2\}$ and $G_2 = \{3\}$. In the first step of the group screening procedure System 1 survives and System 2 is eliminated, as before. System 1 again receives a total sample size of 139 observations. The value of W_{31} then becomes

$$W_{31} = 2.68\left(\frac{10}{10} + \frac{10}{139}\right)^{1/2} \approx 2.77.$$

Thus, System 3 is screened out because

$$\overline{X}_3^{(1)} = 1 \not> 3 - (2.77 - 1)^+ = 1.23.$$

In this case we save the 129 second-stage observations that are not required for System 3. Of course, the results change if the systems are encountered in a different order; in fact, group screening is less efficient than screening all systems at once if $G_1 = \{2, 3\}$ and $G_2 = \{1\}$. We explore this trade-off more fully in §6.4.4.

6. EMPIRICAL EVALUATION

In this section we summarize the results of an extensive empirical evaluation of two of the screening procedures

described above: The combined procedure in §4—which uses screening to deliver a random-sized subset to Rinott's IZ selection procedure—and the group-screening version of this same procedure as presented in §5. We also report one set of results from a smaller study that combined the RSSP of Sullivan and Wilson (1989) with the IZ selection procedure of Dudewicz and Dalal (1975). Recall that an RSSP controls the maximum number of systems that survive screening at the cost of an additional stage of sampling. These procedures are based on weighted sample means; such procedures tend to be more statistically efficient than procedures based on standard sample means, but are also more difficult to justify to practitioners.

Rather than use systems simulation examples, which offer less control over the factors that affect the performance of a procedure, we chose to represent the systems as various configurations of k normal distributions (to assess the impact of departures from normality, we also represented the systems as configurations of k lognormal distributions). In all cases System 1 was the best (had the largest true mean). We evaluated the screening procedures on different variations of the systems, examining factors, including the practically significant difference δ, the initial sample size n_0, the number of systems k, the configuration of the means μ_i, and the configuration of the variances σ_i^2. As a basis for comparison, we also ran Rinott's IZ selection procedure without any screening. The configurations, the experiment design, and the results are described below.

6.1. Configurations

To examine a "difficult" scenario for the screening procedures, we used the slippage configuration (SC) of the means.[3] In the SC, the mean of the best system was set exactly δ or a multiple of δ above the other systems, and all of the inferior systems had the same mean.

To investigate the effectiveness of the screening procedure in removing noncompetitive systems, monotone-decreasing means (MDM) were also used. In the MDM configuration, the means of all systems were spaced evenly apart.

Group-decreasing means (GDM) formed the final configuration. Here the inferior systems were divided into two groups, with means common within groups, but with the first group's mean larger than the second group's. The percentage of systems in each group and the difference between the second group's mean and the best system's mean changed as described below.

In some cases the variances of all systems were equal ($\sigma_i^2 = 1$); in others they changed as described below.

6.2. Experiment Design

For each configuration, we performed 500 macroreplications (complete repetitions) of the entire screening-and-selection procedure. The number of macroreplications was chosen to allow comparison of the true probability of correct selection (PCS) to the nominal level. In all experiments, the nominal PCS was $1 - \alpha = 0.95$ and we took $\alpha_0 = \alpha_1 = 0.025$. If the procedure's true PCS is close to the nominal level, then the standard error of the estimated PCS, based on 500 macroreplications, is about $\sqrt{0.95(0.05)/500} \approx 0.0097$. What we want to examine is how close to $1 - \alpha$ we get. If PCS $\gg 1 - \alpha$ for all configurations of the means, then the procedure is overly conservative.

In preliminary experiments, the first-stage sample size was varied over $n_0 = 5, 10$, or 30. Based on the results of these experiments, follow-up experiments used $n_0 = 10$. The number of systems in each experiment varied over $k = 2, 5, 10, 20, 25, 50, 100, 500$.

The practically significant difference was set to $\delta = d\sigma_1/\sqrt{n_0}$, where σ_1^2 is the variance of an observation from the best system. Thus, we made δ independent of the first-stage sample size by making it a multiple of the standard deviation of the first-stage sample mean. In preliminary experiments, the value of d was $d = 1/2, 1$, or 2. Based on the results of the preliminary experiments, subsequent experiments were performed with $d = 1$ standard deviation of the first-stage sample mean.

In the SC configuration, μ_1 was set as a multiple of δ, while all of the inferior systems had mean 0. In the MDM configuration, the means of systems were spaced according to the following formula: $\mu_i = \mu_1 - b(i-1)$, for $i = 2, 3, \ldots, k$, where $b = \delta/\tau$. Values of τ considered in preliminary experiments were $\tau = 1, 2$, or 3 (effectively spacing each mean δ, $\delta/2$, or $\delta/3$ from the previous mean). For later experiments, the value $\tau = 2$ was used.

In the GDM configuration, the experimental factors considered were the fraction of systems in the first group of inferior systems, η, and the common mean for each group. The fraction in the first group was examined at levels of $\eta = 0.25, 0.5, 0.75$. The means in the first group were all $\mu_i = \mu_1 - \delta$, while the means in the second group were all $\mu_i = \mu_1 - \gamma\delta$. The spacing of the second group was varied according to $\gamma = 2, 3, 4$.

The majority of the experiments were executed with the mean of the best system δ from the next-best system. However, to examine the effectiveness of the screening procedure when the best system was clearly better, some experiments were run with the mean distance as much as 4δ greater. On the other hand, in some cases of the MDM and GDM configurations the mean of the best was less than δ from the next-best system.

For each configuration we examined the effect of unequal variances on the procedures. The variance of the best system was set both higher and lower than the variances of the other systems. In the SC, $\sigma_1^2 = \rho\sigma^2$, with $\rho = 0.5, 2$, where σ^2 is the common variance of the inferior systems. In the MDM and GDM configurations, experiments were run with the variance directly proportional to the mean of each system, and inversely proportional to the mean of each system. Specifically, $\sigma_i^2 = |\mu_i - \delta| + 1$ to examine the effect of increasing variance as the mean decreases, and $\sigma_i^2 = 1/(|\mu_i - \delta| + 1)$ to examine the effect of decreasing

variance as the mean decreases. In addition, some experiments were run with means in the SC, but with variances of all systems either monotonically decreasing or monotonically increasing as in the MDM configuration.

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.

In evaluating the group-screening procedure, we also considered the additional experimental factors of group size, g, and the placement of the best system. The sizes of groups considered were $g = 2, k/2$. Experiments were run with the best system in the first group and in the last group.

When employing restricted-subset selection for screening, we fixed the maximum subset size r to be the larger of $r = 2$ and $r = k/10$ (that is, 10%) of the total number of systems, k. This seemed to be a reasonable imitation of what might be done in practice, because a practitioner is likely to make the subset size some arbitrary but small fraction of k.

6.3. Summary of Results

Before presenting any specifics, we briefly summarize what was observed from the entire empirical study.

The basic combined procedure performed very well in some situations, whereas in others it did not offer significant improvement over Rinott's procedure without screening. In configurations such as MDM, the screening procedure was able to eliminate noncompetitive systems and reduce the total sample size dramatically. However, in the SC, when the practically significant difference and true difference between the best and other systems was less than two standard deviations of the first-stage sample mean, the procedure eliminated few alternatives in the first stage. The key insight, which is not surprising, is that screening is very effective at eliminating systems that are clearly statistically different at the first stage, but is unable to make fine distinctions. Thus, the combined procedure is most useful when faced with a large number of systems that are heterogeneous in performance, not a small number of very close competitors. Restricting the subset size could be more or less efficient than not restricting it, depending on how well the restricted-subset size was chosen.

The PCS can be close to the nominal level $1 - \alpha$ in some situations, but nearly 1 in others. Fortunately, those situations in which PCS is close to 1 were also typically situations in which the procedures were very effective at eliminating inferior systems and thus saved significantly on sampling.

In the robustness study, we found the PCS of the basic combined procedure to be robust to mild departures from normality. However, more extreme departures did lead to significant degradation in PCS, sometimes well below $1 - \alpha$.

The performance of the group-screening procedure was very sensitive to how early a good (or in our experiments, the best) system was encountered. Encountering a good alternative in an early group resulted in savings, while encountering it late (say, in the last group of systems examined) substantially increased the total sampling required relative to screening all systems together. Thus, if groups can be formed so that suspected good candidates are in the first group then this approach can be effective. Boesel et al. (2001) show that it is statistically valid to sort the first-stage sample means before forming the groups, which makes early detection of a good system more likely.

6.4. Some Specific Results

We do not attempt to present comprehensive results from such a large simulation study. Instead, we present details of some typical examples. The performance measures that we estimated in each experiment include the probability of correct selection (PCS), the average number of samples per system (ANS), and the percentage of systems that received second-stage sampling (PSS). Notice that PSS is a measure of the effectiveness of the screening procedure in eliminating inferior systems. We first examine the basic combined procedure from §4, then compare it to the group-screening procedure from §5. We also compare the basic combined procedure to a procedure that employs an RSSP for screening and an IZ selection procedure based on weighted sample means for selection.

6.4.1. Comparisons Among Configurations of the Means. Comparing the performance of the basic combined procedure on all of the configurations of the means shows the areas of strength and weakness. Table 1 provides an illustration. The estimated ANS depends greatly on the configuration of the systems. In the worst case (the SC), the procedure obtains a large number of samples from each system; the ANS grows as the number of systems increases, and it is less efficient than simply applying Rinott's procedure without screening. However, for the MDM the procedure obtains fewer samples per system as the number of systems increases because the additional inferior systems are farther and farther away from the best for MDM. The PSS values indicate that the procedure is able to screen out many systems in the first stage for the MDM configuration, thus reducing the overall ANS. These and other results lead us to conclude that the screening procedure is ineffective for systems within 2δ from the best system (when the δ is exactly one standard deviation of the first-stage sample mean), but quite effective for systems more than 2δ away.

The three configurations also indicate PCS values can be close to or far from the nominal level. For instance, PCS is almost precisely 0.95 for $k = 2$ in the SC, but nearly 1 for most cases of MDM.

The results in Table 2 illustrate the advantages and disadvantages of using a restricted-subset-selection procedure for screening, and also the gain in efficiency from using a procedure based on weighted means.

Comparing the first set of results from Table 1 (which is Rinott's procedure without screening) to the first set of

Table 1. Comparisons across all configurations of the means. In all cases $n_0 = 10, \sigma_i^2 = 1$ for all $i, \delta = 1/\sqrt{n_0}$, and nominal PCS = 0.95.

Procedure/ Configuration	Measure	k = 2	5	10	100	500
no screening	PCS	0.948	0.960	0.960	0.974	0.972
SC	ANS	69	135	184	356	554
$\mu_1 = \delta$	PSS	100%	100%	100%	100%	100%
screening	PCS	0.960	0.964	0.984	0.978	0.968
SC	ANS	86	170	225	453	538
$\mu_1 = \delta$	PSS	89%	91%	94%	99%	99%
screening	PCS	1.000	0.998	0.960	1.000	1.000
MDM	ANS	88	166	177	66	27
$\tau = 2$	PSS	90%	88%	70%	12%	3%
screening	PCS		0.960	0.982	0.994	0.974
GDM	ANS		128	173	393	505
$\eta = 1/2, \gamma = 4$	PSS		65%	69%	80%	89%

results from Table 2 (which is Dudewicz and Dalal's procedure without screening) shows that Dudewicz and Dalal's two-stage indifference-zone procedure is always at least as efficient, in terms of ANS, as Rinott's procedure.[4] Rinott (1978) proved that this is true in general.

When we compare the MDM results from Table 1 with the second set of results in Table 2, we see that restricting the maximum subset size can be either more or less efficient than taking a random-sized subset after the first stage of sampling. In this example, there are exactly 2 systems within δ of the best (the best itself and the second-best system). Therefore, a subset size of $r = 2$ is ideal, making restricted-subset screening more efficient than random-size subset screening when $k = 5$ or 10, because $r = 2$. However, in the $k = 100, 500$ cases a maximum subset of size $r = k/10$ is larger than necessary; thus, the random-size subset screening is substantially more efficient in these cases.

6.4.2. Other Factors. In this section, we look briefly at the effect of varying the number of systems, the practically significant difference, the initial sample size, and the systems' variances.

Increasing the number of systems, k, causes an approximately linear increase in the ANS for the SC (as we saw in Table 1, ANS can decrease when the means are widely spaced as in MDM); see Table 3. For example, with $\delta = \mu_1 = 1/\sqrt{n_0}$ in the SC (and all other systems having mean 0), an increase from 2 systems to 500 systems causes the estimated ANS to increase from 86 to 538.

The practically significant difference δ had a greater effect on the ANS (roughly proportional to $1/\delta^2$) than did k (roughly linear), again with the most significant effect in the SC. For example (see Table 3), at $k = 500$ systems with $\delta = \mu_1 = 1/\sqrt{n_0}$ the ANS = 538, but with $\delta = \mu_1 = 2/\sqrt{n_0}$ the ANS = 134. It is worth noting that linking the δ and the true difference $\mu_1 = \mu_{[k]} - \mu_{[k-1]}$, as we did here, hampers the screening procedure. If instead we *fix* δ and increase only $\mu_{[k]} - \mu_{[k-1]}$, then screening is more effective. Table 4 shows such results when δ is fixed at $1/\sqrt{n_0}$ and the gap between $\mu_{[k]}$ and $\mu_{[k-1]}$ goes from one to four times this amount.

The estimated PCS varied widely with δ, again indicating that the estimated PCS could be close to the nominal value of 95%, ranging from 94.8% to 98.4% in the SC configuration.

The initial sample size, n_0, also had a significant effect on ANS and PCS, particularly in the SC (see Table 5). For instance, with $k = 100$ systems, ANS is greater than 750 for $n_0 = 5$ and $n_0 = 30$, but for $n_0 = 10$ the ANS = 453. The conclusion is that we would like to have a large enough

Table 2. Results for Dudewicz and Dalal, and Sullivan and Wilson combined with Dudewicz and Dalal, for the MDM configuration with $\tau = 2, n_0 = 10, \sigma_i^2 = 1$ for all $i, \delta = 1/\sqrt{n_0}$, and nominal PCS = 0.95.

Procedure	Measure	k = 2	5	10	100	500
no screening	PCS	1.000	1.000	0.998	1.000	1.000
$r = 1$	ANS	69	128	165	312	457
(D&D)	PSS	100%	100%	100%	100%	100%
screening	PCS		1.000	0.998	1.000	1.000
$r = \max\{2, k/10\}$	ANS		106	144	175	209
(S&W + D&D)	PSS		40%	20%	10%	10%

Table 3. The effect of number of systems k and the difference between the best and next-best system for the SC. In all cases $n_0 = 10$, $\sigma_i^2 = 1$ for all i, $\mu_1 = \delta$, $\mu_2 = \mu_3 = \cdots = \mu_k = 0$, where δ is measured in units of $1/\sqrt{n_0}$, and nominal PCS = 0.95.

		k				
δ/Procedure	Measure	2	5	10	100	500
1/2	PCS	0.952	0.966	0.976	0.984	0.964
screening	ANS	362	715	925	1820	2155
	PSS	94%	97%	98%	99%	100%
1	PCS	0.960	0.964	0.984	0.978	0.968
screening	ANS	86	170	225	453	538
	PSS	89%	91%	94%	99%	99%
1	PCS	0.948	0.960	0.960	0.974	0.972
no screening	ANS	69	135	184	356	554
	PSS	100%	100%	100%	100%	100%
2	PCS	0.970	0.982	0.984	0.978	0.968
screening	ANS	16	34	48	109	134
	PSS	64%	64%	70%	91%	97%

initial sample to obtain some sharpness for the screening procedure, while not taking so many observations that we have more precision than really necessary.

The effect of unequal variances was insignificant compared to other experimental factors in the cases we considered, so we limit ourselves to a few comments: When the variance of the best system was increased and the variances of the other systems held constant and equal, then ANS also increased in the SC. However, if all variances were unequal, then as the variance of the best system became larger than those of the inferior systems, the ANS decreased in all configurations. The effect on PCS was not consistent with changes in variances. Overall, the values of the means were more important than the values of the variances, showing that our procedure sacrifices little to accommodate unequal variances.

6.4.3. Robustness. To assess the impact of nonnormal data on the basic combined procedure, the procedure was applied to lognormally distributed data with increasing levels of skewness and kurtosis, relative to the normal distribution (which has skewness 0 and kurtosis 3). Parameters of the lognormal distribution were chosen to obtain the desired variance, skewness, and kurtosis, then the distribution was shifted to place the means in the SC.

Table 6 shows the estimated PCS for three lognormal cases, with the corresponding normal case included for comparison. When skewness and kurtosis differ somewhat from normality (1.780, 9.112), the procedure still maintains a PCS $\geqslant 0.95$. However, as the departure becomes more dramatic, the achieved PCS drops well below the nominal level. A larger number of systems ($k = 100$ vs. $k = 10$) exacerbates the problem, and the degradation is not alleviated by increased sampling (which occurs when δ is smaller). 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.

6.4.4. Effectiveness of Group Screening. In the second portion of the empirical evaluation, the basic combined procedure was compared to a procedure that separated systems

Table 4. The effect of number of systems k and the difference between the best and next-best system for the SC with δ fixed. In all cases $n_0 = 10$, $\sigma_i^2 = 1$ for all i, $\delta = 1/\sqrt{n_0}$, $\mu_2 = \mu_3 = \cdots = \mu_k = 0$, μ_1 is measured in units of $1/\sqrt{n_0}$, and nominal PCS = 0.95.

		k				
μ_1/Procedure	Measure	2	5	10	100	500
1	PCS	0.960	0.964	0.984	0.978	0.988
screening	ANS	86	170	225	453	561
	PSS	89%	91%	94%	99%	99%
1	PCS	0.948	0.960	0.960	0.974	0.972
no screening	ANS	69	135	184	356	554
	PSS	100%	100%	100%	100%	100%
2	PCS	0.998	0.998	1.000	1.000	1.000
screening	ANS	64	155	221	451	537
	PSS	77%	82%	89%	98%	99%
4	PCS	1.000	1.000	1.000	1.000	1.000
screening	ANS	22	89	147	422	532
	PSS	55%	48%	58%	88%	97%

Table 5. The effect of number of systems k and the first-stage sample size n_0 for the SC. In all cases $\sigma_i^2 = 1$ for all i, $\mu_1 = \delta = 1/\sqrt{n_0}$, $\mu_2 = \mu_3 = \cdots = \mu_k = 0$, and nominal PCS = 0.95.

		k				
n_0	Measure	2	5	10	100	500
5	PCS	0.968	0.976	0.968	0.988	0.992
	ANS	69	162	246	765	1651
	PSS	92%	97%	99%	100%	100%
10	PCS	0.960	0.964	0.984	0.978	0.968
	ANS	86	170	225	453	538
	PSS	89%	91%	94%	99%	99%
30	PCS	0.952	0.958	0.952	0.978	0.980
	ANS	282	365	468	817	1074
	PSS	85%	86%	89%	94%	96%

into groups and performed group screening. The performance of the group-screening procedure was mixed.

In the MDM configuration, the performance of the group-screening procedure depended on the placement of the best system, the group size, and interactions between the two factors. When the distance between the best system mean and the inferior system mean was large enough in the SC, then analogous results were obtained.

In general, if the best system was placed in the first group, then the group-screening procedure outperformed the basic combined procedure. For example, for $k = 10$ systems, ANS = 152 and 154 for group screening (with group sizes 2 and $k/2$, respectively), while ANS = 177 with no group screening and 184 with no screening at all; see Table 7. However, if the best system was placed in the last group, then it was better to screen all systems at once.

If the best system was discovered early, then a smaller group size was better. For instance, with $k = 25$ systems, if group size was $g = 2$ (with the best in the first group), then ANS = 105. But if group size was $g = k/2$ (with the best in the first group), then ANS = 126. However, if the best system was in the last position, a larger group size was better (ANS = 314 compared to ANS = 235).

7. CONCLUSIONS

In this paper we have presented a general methodology and several specific procedures for reducing the sampling effort that is required by a two-stage indifference-zone selection procedure. Our approach is to eliminate or screen out obviously inferior systems after the initial stage of sampling, while still securing the desired overall guarantee of a correct selection. Such procedures preserve the simple structure of IZ selection, while being much more efficient in situations where there are many alternative systems, though some are not really competitive.

We focused on procedures that can use screening after an initial stage of sampling—because of our interest in large numbers of systems with heterogeneous performance—and IZ selection procedures based on sample means—because of our preference for standard sample means over weighted sample means. However, in some situations restricting the maximum subset size and using a procedure based on weighted sample means can be advantageous. As a rough rule of thumb, we use first-stage screening when there are a large number of systems and their means are expected to differ widely, but use restricted-subset screening when a substantial number of close competitors are anticipated. The development of more formal methods for choosing among the two is the subject of ongoing research.

All of the combined procedures presented in this paper are based on the assumption that all systems are simulated independently. However, it is well known in the simulation literature that the use of common random numbers (CRN) to induce dependence across systems can sharpen the comparison of two or more alternatives. In the case of R&S, "sharpening" means reducing the total number of observations required to achieve the desired probability of correct selection. Two IZ selection procedures that exploit CRNs have been derived in Nelson and Matejcik (1995), and we have derived combined procedures based on them. How-

Table 6. The effect of nonnormality on PCS for the SC. In all cases $\sigma_i^2 = 1$ for all i, $\mu_1 = \delta$, $\mu_2 = \mu_3 = \cdots = \mu_k = 0$, $n_0 = 10$, δ is measured in units of $1/\sqrt{n_0}$, and nominal PCS = 0.95.

		$\delta = 1/2$		$\delta = 1$	
Distribution	(Skewness, Kurtosis)	$k = 10$ PCS	$k = 100$ PCS	$k = 10$ PCS	$k = 100$ PCS
normal	(0, 3)	0.976	0.984	0.984	0.978
lognormal	(1.780, 9.112)	0.958	0.966	0.962	0.954
lognormal	(4, 41)	0.850	0.774	0.900	0.814
lognormal	(6.169, 113.224)	0.796	0.562	0.848	0.648

Table 7. The effect of group screening relative to screening all at once in the MDM configuration with $\tau = 2$. In all cases $\sigma_i^2 = 1$ for all i, $n_0 = 10$, $\mu_1 = \delta = 1/\sqrt{n_0}$, and nominal PCS = 0.95.

Scenario		$k = 10$	$k = 25$
$g = 2$ with best in first group	PCS	0.992	0.996
	ANS	152	105
	PSS	56%	28%
$g = 2$ with best in last group	PCS	0.998	1.000
	ANS	233	314
	PSS	99%	100%
$g = k/2$ with best in first group	PCS	0.996	1.000
	ANS	154	126
	PSS	59%	35%
$g = k/2$ with best in last group	PCS	1.000	1.000
	ANS	225	235
	PSS	94%	70%
$g = 1$, no group screening	PCS	0.960	1.000
	ANS	177	129
	PSS	70%	36%
no screening	PCS	1.000	1.000
	ANS	184	264
	PSS	100%	100%

ever, these procedures are not desirable when the number of alternatives is very large, which is the focus of this paper. One procedure is based on the Bonferroni inequality, and will become increasingly conservative as k increases. The other assumes that the variance-covariance matrix of the data across all alternatives satisfies a condition known as sphericity. This is an approximation that works well as long as the true variances do not differ too much from the average variance, and the true correlations between systems do not differ too much from the average correlation. As the number of systems increases, the validity of this approximation becomes more and more suspect.

APPENDIX

The following lemmas are used in proving the validity of the screen-to-the-best procedure and the group-screening procedure. A much more complete proof of Lemma 4, as well as a proof of the validity of the screen-to-the-best procedure under common random numbers, and a proof of our inference for systems in the eliminated set are contained in the online companion to this paper.

LEMMA 2 (BANERJEE 1961). *Let Z be an $N(0, 1)$ random variable that is independent of $Y_1, Y_2, \ldots, Y_k$, which are independent chi-squared random variables, with Y_i having degrees of freedom ν_i. Let $\gamma_1, \gamma_2, \ldots, \gamma_k$ be arbitrary weights such that $\sum_{i=1}^k \gamma_i = 1$ and all $\gamma_i \geq 0$. Then*

$$\Pr\left\{Z^2 \leq \sum_{i=1}^k t_i^2 \gamma_i \frac{Y_i}{\nu_i}\right\} \geq 1 - \alpha,$$

when $t_i = t_{1-\alpha/2, \nu_i}$.

LEMMA 3 (TAMHANE 1977). *Let $V_1, V_2, \ldots, V_k$ be independent random variables, and let $g_j(v_1, v_2, \ldots, v_k)$, $j = 1, 2, \ldots, p$, be nonnegative, real-valued functions, each one nondecreasing in each of its arguments. Then*

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

We are now in a position to prove that $\Pr\{[k] \in I | \mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq 1 - \alpha_0$ for the screen-to-the-best procedure. In the proof we let $\delta_{ij} = \mu_i - \mu_j$.

PROOF. Notice that

$$\begin{aligned}
&\Pr\{[k] \in I | \mu_{[k]} - \mu_{[k-1]} \geq \delta\} \\
&\quad = \Pr\{\overline{X}_{[k]} \geq \overline{X}_{[j]} - (W_{[k][j]} - \delta)^+, \ \forall\ j \neq k | \mu_{[k]} - \mu_{[k-1]} \geq \delta\} \\
&\quad = \Pr\left\{\frac{\overline{X}_{[k]} - \overline{X}_{[j]} - \delta_{[k][j]}}{\left(\frac{\sigma_{[k]}^2 + \sigma_{[j]}^2}{n_0}\right)^{1/2}} \geq -\frac{(W_{[k][j]} - \delta)^+}{\left(\frac{\sigma_{[k]}^2 + \sigma_{[j]}^2}{n_0}\right)^{1/2}} - \frac{\delta_{[k][j]}}{\left(\frac{\sigma_{[k]}^2 + \sigma_{[j]}^2}{n_0}\right)^{1/2}}, \right. \\
&\qquad \left. \forall\ j \neq k | \mu_{[k]} - \mu_{[k-1]} \geq \delta \right\}. \qquad \text{(A1)}
\end{aligned}$$

To simplify notation, let

$$Z_j = \frac{\overline{X}_{[k]} - \overline{X}_{[j]} - \delta_{[k][j]}}{\left(\frac{\sigma_{[k]}^2 + \sigma_{[j]}^2}{n_0}\right)^{1/2}}$$

and let

$$v_j = \left(\frac{\sigma_{[k]}^2 + \sigma_{[j]}^2}{n_0}\right)^{1/2}.$$

Now, by the symmetry of the normal distribution, we can rewrite (A1) as

$$\begin{aligned}
&\Pr\left\{Z_j \leq \frac{(W_{[k][j]} - \delta)^+}{v_j} + \frac{\delta_{[k][j]}}{v_j}, \ \forall\ j \neq k \middle| \mu_{[k]} - \mu_{[k-1]} \geq \delta\right\} \\
&\qquad \geq \Pr\left\{Z_j \leq \frac{W_{[k][j]}}{v_j}, \ \forall\ j \neq k\right\}. \qquad \text{(A2)}
\end{aligned}$$

The inequality leading to (A2) arises because $\delta_{[k][j]} \geq \delta$ under the assumed condition, $(W_{[k][j]} - \delta)^+ + \delta \geq W_{[k][j]}$, and Z_j does not depend on $\mu_1, \mu_2, \ldots, \mu_k$.

To further simplify notation, let

$$Q_j = \frac{W_{[k][j]}}{v_j} = t_{(1-\alpha_0)^{1/(k-1)}, n_0 - 1}\left(\frac{S_{[k]}^2 + S_{[j]}^2}{\sigma_{[k]}^2 + \sigma_{[j]}^2}\right)^{1/2}.$$

We now condition on $S_1^2, S_2^2, \ldots, S_k^2$ and rewrite (A2) as

$$\begin{aligned}
&\mathrm{E}\left[\Pr\{Z_j \leq Q_j, \ \forall\ j \neq k | S_1^2, \ldots, S_k^2\}\right] \\
&\quad \geq \mathrm{E}\left[\prod_{j=1}^{k-1} \Pr\{Z_j \leq Q_j | S_1^2, \ldots, S_k^2\}\right] \\
&\quad \geq \prod_{j=1}^{k-1} \mathrm{E}\left[\Pr\{Z_j \leq Q_j\}\right], \qquad \text{(A3)}
\end{aligned}$$

where the first inequality follows from Slepian's inequality (Tong 1980), since $\mathrm{Cov}[Z_i, Z_j] \geq 0$, and the second inequality follows from Lemma 3, since $\Pr\{Z_j \leq Q_j\}$ is increasing in Q_j, and Q_j is increasing in $S_1^2, \ldots, S_k^2$.

To complete the proof, we attack the individual product terms in (A3). Notice that $Q_j \geq 0$ and Z_j is $N(0,1)$, so we can write

$$\begin{aligned}
\Pr\{Z_j \leq Q_j\} &= \frac{1}{2} + \Pr\{0 \leq Z_j \leq Q_j\} \\
&= \frac{1}{2} + \frac{1}{2}\Pr\{Z_j^2 \leq Q_j^2\} \\
&= \frac{1}{2} + \frac{1}{2}\Pr\left\{Z_j^2 \leq t^2\gamma_1 \frac{S_{[j]}^2}{\sigma_{[j]}^2} + t^2\gamma_2 \frac{S_{[k]}^2}{\sigma_{[k]}^2}\right\} \\
&\geq \frac{1}{2} + \frac{1}{2}\left(1 - 2\left(1 - (1-\alpha_0)^{\frac{1}{k-1}}\right)\right) \\
&= (1-\alpha_0)^{\frac{1}{k-1}}, \qquad \text{(A4)}
\end{aligned}$$

where the inequality in (A4) follows from Lemma 2 with $\gamma_1 = 1 - \gamma_2 = \sigma_{[j]}^2/(\sigma_{[k]}^2 + \sigma_{[j]}^2)$.

Substituting this result into (A3) shows that

$$\Pr\{\mathrm{CS}|\mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq \prod_{j=1}^{k-1}(1-\alpha_0)^{\frac{1}{k-1}} = 1 - \alpha_0. \qquad \square$$

To establish the correct selection guarantee for the group-screening procedure, we first prove the following lemma.

LEMMA 4. *Let ℓ^* be the index such that $[k] \in G_{\ell^*}$, and define $F = G_1 \cup G_2 \cup \cdots \cup G_{\ell^*-1}$ (with $F = \varnothing$ if $\ell^* = 1$) and $S = G_{\ell^*+1} \cup G_{\ell^*+2} \cup \cdots \cup G_m$ (with $S = \varnothing$ if $\ell^* = m$). Consider the event*

$$\begin{aligned}
\mathscr{C} = \{&\overline{X}_{[k]}^{(1)} \geq \overline{X}_j^{(1)} - (W_{[k]j} - \delta)^+, \ \forall\, j \in G_{\ell^*} \text{ and} \\
&\overline{X}_{[k]}^{(1)} \geq \overline{X}_j^{(2)} - (W_{[k]j} - \delta)^+, \ \forall\, j \in F \text{ and} \\
&\overline{X}_{[k]}^{(2)} \geq \overline{X}_j^{(1)} - (W_{[k]j} - \delta)^+, \ \forall\, j \in S\}.
\end{aligned}$$

Then $\Pr\{\mathscr{C}|\mu_{[k]} - \mu_{[k-1]} \geq \delta\} \geq 1 - 2\alpha_0$.

PROOF. The proof of this lemma is quite long and tedious, so we only provide a sketch here; the complete details can be found in the online companion.

Using the same initial steps as before, we can show that

$$\begin{aligned}
&\Pr\{\mathscr{C}|\mu_{[k]} - \mu_{[k-1]} \geq \delta\} \\
&\quad \geq \Pr\left\{Z_j^{(1,1)} \leq \frac{W_{[k]j}}{v_j^{(1,1)}}, \ \forall\, j \in G_{\ell^*}; Z_j^{(1,2)} \leq \frac{W_{[k]j}}{v_j^{(1,2)}}, \right. \\
&\qquad \left. \forall\, j \in F; Z_j^{(2,1)} \leq \frac{W_{[k]j}}{v_j^{(2,1)}}, \ \forall\, j \in S\right\}, \qquad \text{(A5)}
\end{aligned}$$

where

$$v_j^{(1,1)} = \left(\frac{\sigma_{[k]}^2}{n_0} + \frac{\sigma_j^2}{n_0}\right)^{1/2},$$

$$v_j^{(1,2)} = \left(\frac{\sigma_{[k]}^2}{n_0} + \frac{\sigma_j^2}{N_j}\right)^{1/2},$$

$$v_j^{(2,1)} = \left(\frac{\sigma_{[k]}^2}{N_{[k]}} + \frac{\sigma_j^2}{n_0}\right)^{1/2},$$

and

$$Z_j^{(a,b)} = \frac{\overline{X}_{[k]}^{(a)} - \overline{X}_j^{(b)} - \delta_{[k]j}}{v_j^{(a,b)}}.$$

The proof then proceeds as follows:

1. Condition on $S_1^2, \ldots, S_k^2$ and use Slepian's inequality to break (A5) into the product of three probabilities.

2. Use Slepian's inequality and Lemma 3 on each piece to break any joint probabilities into products of marginal probabilities for the $Z_j^{(a,b)}$. Altogether, there will be $k-1$ terms in the product.

3. Show that each marginal probability is $\geq (1-2\alpha_0)^{\frac{1}{k-1}}$. It is critical to the proof that W_{ij} is based only on the first-stage sample variances, S_i^2, because S_i^2 is independent of both the first- and second-stage sample means. In the online companion to this paper we provide compelling evidence, but not a proof, that each marginal probability is in fact $\geq (1-\alpha_0)^{\frac{1}{k-1}}$. $\square$

We are now in a position to establish the probability requirement for the group-screening procedure. We claim that

$$\begin{aligned}
&\Pr\{\mathrm{CS}|\mu_{[k]} - \mu_{[k-1]} \geq \delta\} \\
&\qquad \geq \Pr\{\mathscr{C} \text{ and } \overline{X}_{[k]}^{(2)} > \overline{X}_{[j]}^{(2)}, \ \forall\, j \neq k | \mu_{[k]} - \mu_{[k-1]} \geq \delta\}.
\end{aligned}$$

First, the event $\mathscr{C}$ implies that system $[k]$ survives screening the first time it is evaluated and that it is not eliminated by the first-stage sample means of any systems evaluated after $[k]$. This event occurs with probability $\geq 1 - 2\alpha_0$ by Lemma 4. Then the event $\{\overline{X}_{[k]}^{(2)} > \overline{X}_{[j]}^{(2)}, \ \forall\, j \neq k\}$ insures that the second-stage sample mean of system $[k]$ is not eliminated by any other system's second-stage sample mean; this event occurs with probability $\geq 1 - \alpha_1$ because we used Rinott's constant to determine the second-stage sample size. Therefore, using Lemma 1, the probability of the joint event is $\geq 1 - 2\alpha_0 - \alpha_1$. $\square$

ENDNOTES

[1] However, it can be shown that indifference-zone selection procedures are quite efficient when the worst-case really happens; see for instance Hall (1959), Eaton (1967), and Mukhopadhyay and Solanky (1994).

[2] An RSSP insures that the maximum size of the subset does not exceed a user-specified number of systems by using two or more stages of sampling to form the subset.

[3] The slippage configuration is the least favorable configuration for Rinott's procedure, which forms the second stage of two of the procedures presented here. "Least favorable" means that this configuration attains the lower bound on probability of correct selection over *all* configurations of the means with $\mu_{[k]} - \mu_{[k-1]} \geq \delta$.

[4] Notice that in the absence of screening the ANS of Rinott or Dudewicz and Dalal is independent of the configuration of the true means.

ACKNOWLEDGMENT

This research was partially supported by National Science Foundation Grant numbers DMI-9622065 and DMI-9622269, and by JGC Corporation, Symix Corporation, and Rockwell Software. The reports of three referees and an associate editor were very helpful.

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.

Bechhofer, R. E., C. W. Dunnett, D. M. Goldsman, M. Hartmann. 1990. A comparison of the performances of procedures for selecting the normal population having the largest mean when the populations have a common unknown variance. *Comm. Statist.* **B19** 971–1006.

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

Boesel, J., B. L. Nelson. 1998. Accounting for randomness in heuristic simulation optimization. *Proc. 12th European Simulation Multiconference*. Manchester, U.K. Society for Computer Simulation International. 634–638.

——, ——, S. H. Kim. 2001. Using ranking and selection to clean up after a simulation search. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Dudewicz, E. J., S. R. Dalal. 1975. Allocation of observations in ranking and selection with unequal variances. *Sankhyā* **B37** 28–78.

Eaton, M. L. 1967. Some optimum properties of ranking procedures. *Ann. Math. Statist.* **38** 124–137.

Fu, M. 1994. Stochastic optimization via simulation: A review. *Ann. Oper. Res.* **53** 199–248.

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

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

——. 1965. On some multiple decision (selection and ranking) rules. *Technometrics* **7** 225–245.

——, W.-C. Kim. 1984. A two-stage elimination-type procedure for selecting the largest of several normal means with a common unknown variance. T. J. Santner, A. C. Tamhane eds. *Design of Experiments: Ranking and Selection—Essays in Honor of Robert E. Bechhofer* Marcel Dekker, New York, 77–94.

Hall, W. J. 1959. The most economical character of Bechhofer and Sobel decision rules. *Ann. Math. Statist.* **30** 964–969.

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 Anal.* **10** 1–16.

Hochberg, Y., R. Marcus. 1981. Three stage elimination type procedures for selecting the best normal population when variances are unknown. *Comm. Statist.* **A10** 597–612.

——, A. C. Tamhane. 1987. *Multiple Comparison Procedures*. John Wiley, New York.

Jacobson, S., L. Schruben. 1989. A review of techniques for simulation optimization. *Oper. Res. Letters* **8** 1–9.

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

Koenig, L. W., 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. *Comm. Statist.* **B14** 719–734.

Mukhopadhyay, N., T. K. S. Solanky. 1994. *Multistage Selection and Ranking Procedures: Second-order Asymptotics*. Marcel Dekker, New York.

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

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

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

Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Comm. Statist.* **A7** 799–811.

Santner, T. J. 1975. A restricted subset selection approach to ranking and selection problems. *Ann. Statist.* **3** 334–349.

——, M. Behaxeteguy. 1992. A two-stage procedure for selecting the largest normal mean whose first stage selects a bounded random number of populations. *J. Statist. Planning and Inference* **31** 147–168.

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

Tamhane, A. C. 1976. A three-stage elimination type procedure for selecting the largest normal mean (common unknown variance). *Sankhyā* **B38** 339–349.

——. 1977. Multiple comparisons in model I one-way anova with unequal variances. *Comm. Statist.* **A6** 15–32.

——. 1980. On a class of multistage selection procedures with screening for the normal means problem. *Sankhyā* **B42** 197–216.

——, R. E. Bechhofer. 1977. A two-stage minimax procedure with screening for selecting the largest normal mean. *Comm. Statist.* **A6** 1003–1033.

——, ——. 1979. A two-stage minimax procedure with screening for selecting the largest normal mean (ii): an improved PCS lower bound and associated tables. *Comm. Statist.* **A8** 337–358.

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

Wilcox, R. R. 1984. A table for Rinott's selection procedure. *J. Quality Tech.* **16** 97–100.