Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

ADAPTIVELY CHOOSING THE BEST PROCEDURE FOR SELECTING THE BEST SYSTEM

Justin Boesel

The MITRE Corporation
1820 Dolley Madison Boulevard
McLean, VA 22102, U.S.A.

ABSTRACT

This paper presents a method that uses initial sample data to choose between statistical procedures for identifying the simulated system with the best (maximum or minimum) expected performance. The method chooses the procedure that minimizes the additional number of simulation replications required to return a pre-specified probability guarantee. This problem may be encountered *after* a heuristic search procedure has been applied in a simulation-optimization context. In this setting, initial samples from each system may already have been taken, but because of stochastic variation, the system with the best sample mean at the end of the search procedure may not be the true best system encountered during the search. Empirical work in previous papers suggests that the relative number of additional replications required by existing procedures depends on factors — such as the configuration of the systems' means and their variances — that may be unknown prior to initial data collection. These results motivated the approach taken in this paper, where we postpone the choice between statistical procedures until after observing the initial data.

1 INTRODUCTION

In this paper we address the problem of choosing the best statistical procedure for finding the simulated system with the best (maximum or minimum) expected performance when initial samples from each system have already been taken. This situation is likely to be encountered at the end of a heuristic simulation-optimization run, where a search procedure may have uncovered very good solutions, but cannot guarantee which solution is the true best among those visited. In a stochastic setting, an inferior solution may *seem* better than the true best solution — that is, it may have a better sample mean — simply because of stochastic variation. If one takes the solution with the best sample mean at the end of the simulation-optimization run, there is some (unknown) chance that another solution visited by the search is actually better, but that the variance of the output measure has produced misleading results.

Statistical procedures, specifically subset-selection and indifference-zone (IZ) ranking procedures, can help to reduce (or at least bound) the chance that an inferior solution is returned as the best. A single-stage subset-selection procedure, which requires no additional simulation replications, returns a random-sized subset that contains the best of the k systems with probability $\geq 1-\alpha$. Two-stage (IZ) procedures, which require additional sampling of the competitive systems, guarantee to select the best system with probability $\geq 1-\alpha$ whenever the best is at least a user-specified amount, δ, better than the others. If there are some near-best solutions within δ of the best, most two-stage IZ procedures will return the best or one of these near-best solutions. The user-specified quantity, δ, is called the indifference zone, and it represents the smallest difference worth detecting (Bechhofer, Santner and Goldsman 1995).

While helpful, both subset-selection and IZ procedures have shortcomings that hamper their usefulness in a simulation-optimization setting. A single-stage subset-selection procedure requires no additional simulation effort after the search has finished, but it may not eliminate many (or any) systems. On the other hand, an IZ procedure guarantees to return a single system within δ of the best with a pre-specified probability, but it may require an enormous amount of additional simulation effort to do so. In our environment, however, we may have hundreds or thousands of systems to consider, making the simulation effort required to use an IZ procedure alone in such a setting prohibitive. Fortunately, the two approaches (subset and IZ) can work together to deliver a single system, while meeting our indifference and probability requirements with less simulation effort than would be required by the IZ procedure alone.

Several of these combined procedures are discussed in Boesel, Nelson, and Kim (2000) and Nelson et al., (1998). One procedure, which we will call Screen-and-Continue, screens out clearly inferior systems using a subset selection procedure, retains the first-stage simulation data, then employs the second-stage of an IZ procedure, collecting

additional data on the remaining systems, to determine the single best. Another procedure, called Screen-Restart-and-Select, is similar in that is uses a subset selection procedure to eliminate inferior systems, but then discards the first-stage data, performing an independent two-stage IZ procedure on the remaining systems.

Empirical work in Boesel, Nelson and Kim (2000) suggests that the *relative* number of replications required by each of these procedures depends upon a number of factors, such as the spacing of the systems' means, the systems' variances, and the number of initial replications taken at each system. In some situations, it is better to use Screen-and-Continue, while in other situations, Screen-Restart-and-Select is superior. Of course, without any data, it is nearly impossible to determine which procedure will require fewer additional replications. Observing the initial sample data, however, can give a clearer picture about these factors, making it easier to choose between the two procedures.

The current article describes a method for choosing between two such combined procedures *after* observing the initial sample data and the results of the screening procedure. While this method may sound questionable, in Section 5 we prove that this method is statistically valid, although the provable probability of correct selection is somewhat degraded.

The remainder of this paper is divided into five sections. Section 2 provides notation and descriptions of the Screen-and-Continue and the Screen-Restart-and-Select procedures. Section 3 describes the Choice procedure, which uses a decision rule to choose between the Screen-and-Continue and the Screen-Restart-and-Select procedures *after* observing the first-stage data. Section 4 presents the results of an empirical study comparing the Choice procedure with its component procedures — Screen-and-Continue and Screen-Restart-and-Select — in a variety of settings. Section 5 provides a lower bound on the probability of correct selection under the Choice procedure, and Section 6 draws some conclusions.

2 BACKGROUND

2.1 Notation and Assumptions

We assume that a preliminary or *first-stage* set of simulation output data generated by a search procedure is "dropped into our laps." Let k be the number of different systems in the data set, and let n_0 be the number of replications already performed on each system. Further, let X_{im} be the output from replication m of system i, which we assume are i.i.d. $\mathrm{N}(\mu_i, \sigma_i^2)$ random variables. Systems are to be compared based on their true means, μ_i, and we assume that larger μ_i is better throughout this paper. The first-stage sample mean of system i is $\bar{X}_i^{(1)}$, while the overall, two-stage sample mean of system i is $\bar{X}_i^{(2)}$. The initial first-stage sample variance of system i, used in the screening phase of both procedures, is S_{0i}^2. The Screen-Restart-and-Select procedure also makes use of the restarted first-stage sample variance, denoted by S_{ri}^2, based on a sample of size n_r. For clarity's sake, we will assume that the number of first-stage replications taken is equal across systems. Boesel, Nelson, and Kim, (2000) show how this assumption can be relaxed.

2.2 Screen-and-Continue Procedure

Nelson et al. (1998) developed a provably valid Screen-and-Continue procedure that retains the original first-stage sample data after screening. A description of this procedure, which we also will refer to as Continuation, follows.

Screen-and-Continue Procedure

1. Sample $X_{im}, i = 1, 2, \ldots, k,\ m = 1, 2, \ldots, n_0$, where the X_{im} are i.i.d. $\mathrm{N}(\mu_i, \sigma_i^2)$ random variables.
2. Select the desired confidence level, $1 - \alpha$, and the indifference level, δ.
3. Run the subset procedure (described below). To obtain an overall confidence level of $1 - \alpha$, we set $1 - \alpha_0 = \sqrt{1 - \alpha}$ for the screening procedure and $1 - \alpha_1 = \sqrt{1 - \alpha}$ for the selection procedure; however, any decomposition whose product is $1 - \alpha$ could be used.
   (a) Let
   $$W_{ij} = \frac{t}{\sqrt{n_0}} \left(S_{0i}^2 + S_{0j}^2 \right)^{\frac{1}{2}} \qquad (1)$$
   where $t = t_{(1-\alpha_0)^{\frac{1}{k-1}}, n_0 - 1}$.
   (b) Set
   $$I = \left\{ i : 1 \le i \le k \text{ and } \bar{X}_i^{(1)} \ge \bar{X}_j^{(1)} - W_{ij}, \forall j \ne i \right\}.$$
   (c) Return I, the group of systems that survive the screen and let $M = |I|$.
4. Calculate the total required sample size from system $i \in I$, N_i, as
   $$N_i = \max \left\{ n_0, \left\lceil \left(\frac{h_k S_{0i}}{\delta} \right)^2 \right\rceil \right\} \qquad (2)$$
   where $h_k = h(k, (1 - \alpha_1), n_0)$ is Rinott's (1978) constant where the number of systems being compared is k, the confidence level is $(1 - \alpha_1)$, and

the first-stage sample size is n_0. Let $\lceil a \rceil$ denote the smallest integer greater than a.

5. Take $N_i - n_0$ additional observations from each system $i \in I$.
6. Of the M surviving systems, select as best the system i with the largest overall sample mean $\bar{X}_i^{(2)} = \sum_{m=1}^{N_i} X_{im}/N_i$.

Unfortunately, the validity guarantee for this procedure requires that the critical value h_k used in the IZ procedure be determined as though all k systems remain in contention, rather than just the M that survive screening. This is because the procedure uses the initial samples from the search in the IZ procedure (this may not be obvious, but the conditional probability of selecting the best system, given it passed screening, depends upon whether or not the first-stage data are retained). Thus, h_k remains large, so N_i is also large.

If, however, we re-run the first-stage samples of the M systems that survive screening, we can eliminate some of these problems. Restarting allows us to use M, rather than the original k, in our determination of Rinott's constant. This could reduce the constant, perhaps dramatically. In many cases, the savings gained through this reduction from h_k to h_M more than offsets the losses involved in re-running the first-stage samples.

2.3 Screen-Restart-and-Select Procedure

The combined procedure presented below is simple and statistically valid; it employs a subset-selection procedure to screen out inferior systems, then discards the original data and employs an *independent* two-stage IZ procedure on the survivors by taking a new first-stage sample from each. We will refer to this procedure as Restart.

Screen-Restart-and-Select Procedure

1-3. Same as under Screen-and-Continue.

4. Take independent samples of size $n_r \geq 2$ from each system $i \in I$ (discarding the initial first-stage sample), and calculate a new sample variance estimate, S_{ri}^2, from the new sample.
5. Calculate the total required sample size from system $i \in I$, N_i, as

$$N_i = \max\left\{n_r, \left\lceil \left(\frac{h_M S_{ri}}{\delta}\right)^2 \right\rceil\right\} \qquad (3)$$

where $h_M = h(M, (1-\alpha_1), n_r)$ is Rinott's (1978) constant where the number of systems being compared is M, the confidence level is $(1-\alpha_1)$, and the first-stage sample size is n_r.

6. Take $N_i - n_r$ additional observations from each system $i \in I$.
7. Of the M surviving systems, select as best the system i with the largest overall sample mean $\bar{X}_i^{(2)} = \sum_{m=1}^{N_i} X_{im}/N_i$.

Although the Restart procedure has easily provable statistical properties, it is unfortunate that it discards data. If the initial sample size is large or if the screen fails to eliminate many systems, re-running the initial samples becomes wasteful.

3 THE CHOICE PROCEDURE

The tradeoffs between the Continuation and Restart procedures are fairly straightforward: if screening is effective, eliminating a large number of systems, then the benefit of Restart will be great, because Rinott's constant will be greatly reduced ($h_M << h_k$). If, on the other hand, the number of initial replications, n_0, is large, or if screening does not eliminate many systems, then the Continuation procedure may fare better.

The effectiveness of screening depends not only upon n_0, but also upon the configuration of the systems (the spacing of their means), and the within-system variance of each system. These factors are impossible to observe without observing the first-stage data.

Boesel, Nelson, and Kim (2000) conduct an empirical study that compares the Restart and Continuation procedures in a variety of settings, with different configurations, variances and initial sample sized. Neither procedure dominated the other in terms of the number of additional replications required to return a statistical guarantee.

We consider an approach that postpones the choice between the Restart and Continuation Procedures until *after* we have observed the first-stage data. Under this *Choice* procedure, Restart is chosen only if it results in a smaller (estimated) total expected number of replications than does Continuation. The choice boils down to the following: does the reduction in h (due to screening out systems) under Restart make up for the cost of re-running the initial samples for the survivors? Although this procedure sounds questionable, in Section 5 we prove that $\Pr\{CS\} \geq 1-3\alpha/2$, where "CS" is the event of correctly selecting the best system.

The rule used to decide which procedure to employ after viewing the first-stage data is simple; choose the procedure with the lower number of additional required replications. The total number of additional replications required under the Continuation procedure is

$$\sum_{i \in I}\left(\max\left\{n_0, \left\lceil \left(\frac{h_k S_{0i}}{\delta}\right)^2 \right\rceil\right\}\right) - Mn_0.$$

The *actual* number of additional replications required under the Restart procedure cannot be calculated immediately after screening, but must be estimated. The *expected* number of additional replications required is

$$\sum_{i \in I} \max \left\{ n_r, \left\lceil \left(\frac{h_M S_{0i}}{\delta} \right)^2 \right\rceil \right\}.$$

4 EMPIRICAL STUDY

We conducted an extensive empirical evaluation to compare the Continuation, Restart, and Choice procedures introduced in this paper to each other. The systems are represented as various configurations of k normal distributions. We evaluated the procedures on different variations of the systems, examining factors including: the number of systems, k; the number of initial replications, n_0; the within-system variance, σ_i^2; and the configuration of the means, μ_i, for $i = 1, 2, \ldots, k$.

4.1 Experiment Design

In all cases, the best system was system 1 and its true mean was set to 1 ($\mu_1 = 1$). To examine a scenario in which screening was unlikely to eliminate many systems, we used the slippage configuration (SC) of the means. In the SC, the mean of the best system was set exactly one indifference zone, δ, above the other systems, and all of the inferior systems had the same mean. To investigate a setting in which screening was likely to eliminate many systems, we also used monotone decreasing means (MDM). In the MDM configuration, the means of all systems were spaced evenly apart. The size of the spaces between systems were set at δ. In both the SC and MDM configurations, $\delta = 1$.

For both the Restart and the Continuation procedures where no choice was allowed, we set the nominal probability of correct selection (PCS) to $1-\alpha = 0.95$. (Throughout the remainder of this section, we will refer to these experiments as Restart950 and Continue950, respectively.)

For the Choice procedure, we conducted experiments setting the nominal PCS at both 0.95 and 0.925. (We will refer to these experiments as Choice950 and Choice925, respectively.) We set the PCS of 0.925 to perform the screening and selection *components* at $(1-\alpha_0)$ and $(1-\alpha_1)$, respectively, the same levels used in Restart950 and Continue950. The overall PCS of 0.925 for Choice925 represents the degradation of the nominal PCS due to choosing between Restart and Continuation after viewing the first-stage data (PCS $= 1-3\alpha/2$). We set the nominal PCS for Choice950 at 0.950 to compare it to Restart950 and Continue950. To achieve an overall PCS of 0.95 for Choice950, we reduced α_0 and α_1, the nominal PCS levels of the component screening and selection, to 0.017, so that $\alpha = 0.033$, and PCS $= 1 - 3\alpha/2 = .95$.

In the experiments, 500 macroreplications (complete repetitions of the entire experiment) were performed for each configuration. If the procedure's true PCS is close to the nominal level, then the standard error of the estimated PCS, based on 500 macroreplications and PCS $= 0.950$, is near $\sqrt{0.95(0.05)/500}$, which is approximately 0.0097. For PCS $= 0.925$, the standard error is near $\sqrt{0.925(0.075)/500}$, which is approximately 0.0118. Since we are guaranteed that PCS $\geq 1-\alpha$ for normally distributed data, we want to examine how close to $1-\alpha$ we get. If PCS $\gg 1-\alpha$ for all configurations of the means, then the procedure is overly conservative.

The first-stage sample size varied over $n_0 = 5, 10, 20$ from one experiment to the next. In all experiments, the first-stage sample size under Restart, n_r, was set equal to the initial first-stage sample size, n_0. The true variance varied over $\sigma^2 = 1.0, 5.0$ from one experiment to the next. In each experiment, every system had equal variance. The number of systems varied over $k = 5, 10, 25, 100, 500$. All told, we ran 60 experiments (2 configurations (MDM and SC) $\times$ 2 variance levels $\times$ 3 settings for n_0 $\times$ 5 settings for k).

4.2 Results

Rather than present comprehensive results from such a large simulation study, we point out the main trends and present details of some illustrative 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. For the Choice experiments, we also estimated the percentage of trials in which Restart and Continuation were chosen, as well as the percentage of trials in which screening successfully eliminated all but one system, so neither Restart nor Continuation was necessary.

A number of patterns emerged from the experiments. Most importantly, there were only negligible differences between the *observed* PCS of Restart950 and Continue950, both of which have guaranteed PCS values of 0.95, and Choice925, which has a guaranteed PCS value of 0.925. These results suggest that while our *guaranteed* PCS will be degraded by employing the Choice procedure, the *actual* PCS may not be.

As expected, the ANS value of Choice925 was almost always lower than the ANS value of either Restart950 or Continue950. In several instances, the ANS value of Choice925 was slightly lower than the ANS value of *both* options. In only one case was the ANS value of Choice925 (slightly) higher than the ANS value of both Restart950

and Continue950. These results indicate that the Choice procedure did a good job of choosing between Restart and Continuation. More specifically, it indicates that S^2_{0i}, used by the Choice procedure to *estimate* the number of additional replications required by Restart, was usually an adequate predictor of S^2_{0r}.

Unfortunately, Choice950, which employed component procedures with higher PCS guarantees to return the same overall PCS guarantee as Continue950 and Restart950, did not perform as well as Choice925. In general, Choice950 fared poorly unless one option (Restart or Continuation) had a much lower ANS value than the other. When Restart950 and Continue950 had similar ANS values, Choice950 often required more replications than *both*. Furthermore, despite the additional replications required, the observed PCS values of Choice950 were, by and large, no better than those of Choice925. While this last result is somewhat surprising, it should be noted that the observed PCS values of all of the procedures were, for the most part, quite high, indicating that the procedures are overly conservative. Typically, experiments with systems in the MDM configuration yielded PCS values over 0.99, while experiments with systems in the SC configuration yielded somewhat lower PCS values, usually between 0.97 and 0.98.

By and large, the Choice procedure tended to select Continuation when the systems were configured in the slippage configuration, when the number of systems was small, or when n_0 was large. Not surprisingly, the Choice procedure tended to select Restart when a large number of systems were configured in the MDM configuration, or when n_0 was small.

Tables 1, 2, and 3 present some results from a set of experiments in which the systems are in the MDM configuration, the initial number of replications $n_0 = 10$, and all systems have equal variance of 5.0.

In Table 1, we can see that as the number of systems, k, increases, the percentage of systems surviving screening falls. As this occurs, Table 2 shows that the number of replications required by the Restart procedure falls below the number required by the Continuation procedure. Table 3, which presents the percentage of trials in which the Choice procedures chose Restart over Continuation, shows the impact of the increasingly effective screen. As the number of systems increases, and the percentage of systems surviving screening decreases, the Choice procedures choose Restart more and more frequently. (In all of these trials, more than one system survived screening so either Restart or Continuation was always necessary.)

Table 2 shows that the average number of replications required by the Choice925 procedure, which used the same component procedures as Continue950 and Restart950, was *always* lower than the number required by the greater of Continue950 and Restart950.

Table 1: Percentage of Systems Receiving Second-stage Sampling, by Procedure and Number of Systems, k (MDM, $n_0 = 10$, $\sigma^2 = 5.0$)

	k=5	k=10	k=25	k=100	k=500
Continue950	96%	76%	37%	11%	3%
Restart950	96%	76%	37%	11%	3%
Choice950	98%	79%	40%	12%	3%
Choice925	96%	76%	37%	11%	3%

Table 2: Average Samples per System, by Procedure and Number of Systems, k (MDM, $n_0 = 10$, $\sigma^2 = 5.0$)

	k=5	k=10	k=25	k=100	k=500
Continue950	88.1	98.8	72.6	37.1	17.5
Restart950	94.2	89.4	52.9	23.8	13.6
Choice950	101.1	109.8	62.8	26.7	14.3
Choice925	87.7	92.4	52.9	23.8	13.6

Table 3: Percentage of Trials in which the Choice Procedure Selected Restart over Continuation, by Guaranteed PCS and Number of Systems, k (MDM, $n_0 = 10$, $\sigma^2 = 5.0$)

	k=5	k=10	k=25	k=100	k=500
Choice950	4%	60%	100%	100%	100%
Choice925	3%	55%	100%	100%	100%

To illustrate the patterns under the slippage configuration, Tables 4, 5, and 6 present some results from a set of experiments in which the systems are in the SC, the initial number of replications $n_0 = 5$, and all systems have equal variance of 1.0.

In the slippage configuration, screening is very difficult, and Table 4 shows that screening eliminated very few systems in this set of experiments. Table 5 shows the results of this weak screening: Restart, which depends on screening to lower Rinott's constant, h, did not lower ANS a great deal in this setting, despite that fact that the initial number of replications — and the penalty involved with discarding them — was low ($n_0 = 5$). Table 6 shows that, because Restart was not particularly helpful, the Choice procedures did not select Restart as frequently as when systems were in the MDM configuration.

The experiments also yielded some noteworthy direct comparisons of Restart950 and Continue950. Restart fared worst relative to Continuation when systems were in the slippage configuration (SC) and variance was high, making screening ineffective. In these situations, Restart essentially throws away all n_0 initial replications from all k systems for no benefit. In our experiments, the ANS value of Restart950 only exceeded that of Continue950 by more than n_0 when n_0 was low (5 or 10), variance was high ($\sigma = 5.0$), *and* the systems were in the slippage configuration. Even in these instances, the differences were never greater than $2n_0$. On the other hand, the amount by which Continuation's ANS exceeded Restart's ANS was much greater. Continuation

Table 4: Percentage of Systems Receiving Second-stage Sampling, by Procedure and Number of Systems, k (SC, $n_0 = 5$, $\sigma^2 = 1.0$)

	k=5	k=10	k=25	k=100	k=500
Continue950	92%	93%	95%	97%	98%
Restart950	92%	93%	95%	97%	98%
Choice950	93%	95%	96%	98%	98%
Choice925	92%	93%	95%	97%	98%

Table 5: Average Samples per System, by Procedure and Number of Systems, k (SC, $n_0 = 5$, $\sigma^2 = 1.0$)

	k=5	k=10	k=25	k=100	k=500
Continue950	32.8	50.4	80.6	160.2	480.1
Restart950	35.0	51.3	80.4	158.4	466.2
Choice950	39.5	61.2	99.6	193.8	518.4
Choice925	31.9	49.0	79.2	157.9	467.1

Table 6: Percentage of Trials in which the Choice Procedure Selected Restart over Continuation, by Guaranteed PCS and Number of Systems, k (SC, $n_0 = 5$, $\sigma^2 = 1.0$)

	k=5	k=10	k=25	k=100	k=500
Choice950	18%	14%	11%	14%	57%
Choice925	16%	16%	15%	17%	73%

fared worst relative to Restart when screening was effective; that is, when a large number of systems were in the MDM configuration. In one experiment, where $k = 500$, $n_0 = 5$, $\sigma^2 = 5.0$ and MDM was used, Restart950 had an ANS of 45, while Continue950 had an ANS of 274. This type of scenario, with many widely-spaced systems with just a few replications each, is important because one is likely to encounter it after a heuristic search procedure has concluded.

5 LOWER BOUND ON PCS UNDER CHOICE PROCEDURE

Below, we will prove that under the Choice procedure $\Pr\{CS\} \geq (1-\alpha_0) \times (1-2\alpha_1)$, where $1-\alpha_0$ is the confidence level used in the screening phase, and $1-\alpha_1$ is the confidence level used in the selection phase.

For notation, let B be the event that the best system survives screening, while D_R is the event that the decision rule—whatever it is—favors restart. A "bar" over an event indicates its complement. Let the subscript C indicate probabilities computed under the assumption that we always continue, while R indicates probabilities computed under the assumption that we always restart.

Using this notation, we can write the probability of a correct selection under the Choice procedure, given that the best system survives screening, as

$$\begin{aligned}\Pr\{CS|B\} &= \Pr_C\{CS|B, \bar{D}_R\}\Pr\{\bar{D}_R\} + \\ &\quad \Pr_R\{CS|B, D_R\}\Pr\{D_R\}.\end{aligned}$$

First, we will find a lower bound on $\Pr_C\{CS|B, \bar{D}_R\}\Pr\{\bar{D}_R\}$. We know from Nelson et al. (1998) that the probability of selecting the best in the Continuation procedure, given that the true best survives screening, is greater than or equal to $1-\alpha_1$. In our notation,

$$\Pr_C\{CS|B\} \geq 1-\alpha_1.$$

Conditioning on the outcome of the decision rule yields

$$\begin{aligned}\Pr_C\{CS|B\} &= \Pr_C\{CS|B, D_R\}\Pr\{D_R\} + \\ &\quad \Pr_C\{CS|B, \bar{D}_R\}\Pr\{\bar{D}_R\} \geq 1-\alpha_1.\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr_C\{CS|B, \bar{D}_R\}\Pr\{\bar{D}_R\} &\geq 1-\alpha_1 - \\ &\quad \Pr_C\{CS|B, D_R\}\Pr\{D_R\} \\ &\geq 1-\alpha_1-\Pr\{D_R\}. \qquad (4)\end{aligned}$$

We know that under Restart, if the best survives screening, the probability of success is greater than or equal to $1-\alpha_1$, regardless of the outcome of the decision rule. As a result, $\Pr_R\{CS|B, D_R\} \geq 1-\alpha_1$. Combining this result with (4) yields,

$$\begin{aligned}\Pr\{CS|B\} &= \Pr_C\{CS|B, \bar{D}_R\}\Pr\{\bar{D}_R\} + \\ &\quad \Pr_R\{CS|B, D_R\}\Pr\{D_R\} \\ &\geq 1-\alpha_1-\Pr\{D_R\}+(1-\alpha_1)\Pr\{D_R\} \\ &= 1-\alpha_1-\alpha_1\Pr\{D_R\} \\ &\geq 1-2\alpha_1.\end{aligned}$$

Consequently, the overall Choice procedure (screening and selection phases) yields

$$\Pr\{CS\} \geq (1-\alpha_0)\times(1-2\alpha_1)$$

which is equal to $(1-\alpha/2)\times(1-\alpha)$ if $\alpha_0 = \alpha_1 = \alpha/2$. As a result,

$$\Pr\{CS\} \geq 1-3\alpha/2+\alpha^2/2 \geq 1-3\alpha/2.$$

6 CONCLUSIONS

The Choice procedure presented in this article allows one to view initial sample data and screening results before deciding whether to retain initial sample data and continue with a two-stage selection-of-the-best procedure or to discard the initial data and restart such a procedure. This option can save the amount of simulation effort required to return a pre-specified

PCS guarantee. Unfortunately, this flexibility comes at a price; to effectively hedge against picking the more costly option, the guaranteed PCS of the overall procedure falls from $1-\alpha$ to $1-3\alpha/2$. On the bright side, our experiments suggest that while the guaranteed PCS is degraded, the actual PCS may be unaffected.

Of course, while Continuation may be *much* more costly (in terms of required replications) than Restart, Restart, at worst, is only somewhat more costly than Continuation. So, from a practical standpoint, using Restart without Choice may provide an adequate hedge against excessive cost, especially if n_0 is small. But, if n_0 is large and there is a good chance that screening will be ineffective, Choice may be the less costly option.

ACKNOWLEDGMENTS

The author would like to thank Prof. Barry L. Nelson of Northwestern University for his guidance.

REFERENCES

Bechhofer, R. E., T. J. Santner and D. Goldsman. 1995. *Design and analysis for statistical selection, screening and multiple comparisons*. New York: John Wiley and Sons.

Boesel, J., B. L. Nelson, and S. H. Kim. 2000. Using ranking and selection to 'clean up' after simulation optimization. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 1998. Simple procedures for selecting the best simulated system when the number of alternatives is large. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University.

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

AUTHOR BIOGRAPHY

JUSTIN BOESEL is a Senior Simulation and Modeling Engineer at the MITRE Corporation in McLean, Virginia. He received a B.A. degree in history from Southern Methodist University; and he received M.S. and Ph.D. degrees in Industrial Engineering & Management Sciences from Northwestern University. He is a member of INFORMS. His email address is <boesel@mitre.org>.