Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

METHODS FOR SELECTING THE BEST SYSTEM

David Goldsman

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Barry L. Nelson

Department of Industrial & Systems Engineering
The Ohio State University
Columbus, Ohio 43210

Bruce Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

In this tutorial we consider three methods for selecting the best of a set of competing systems: interactive analysis, ranking and selection, and multiple comparisons. We describe each method; discuss assumptions, implementation aspects, advantages, and disadvantages; and demonstrate the use of each method with an airline-reservation-system simulation example.

1 INTRODUCTION

Stochastic simulation is often used to compare competing systems. In this tutorial, we discuss concepts and methods to select the best system. The goal is to design an efficient experiment and to provide a sound data analysis.

We consider three approaches: interactive analysis (IA), ranking and selection (R&S), and multiple comparison procedures (MCPs). The concept of standard error underlies all three approaches, but they differ in that IA, R&S, and MCPs are based on estimation, optimization, and inference, respectively.

After considering factors affecting problem context, we define a specific example involving the selection of an airline-reservation system. We pursue the example with Schmeiser discussing IA in Section 2, Goldsman discussing R&S in Section 3, and Nelson discussing MCPs in Section 4. Common notation is used throughout.

1.1 The Problem Context

The problem context can vary dramatically. Some factors that affect the choice of design and analysis methods include:

- Number of competing systems, k. In our example k is small, less than ten or twenty, and the systems are given. Another possibility is k large, possibly infinite, as is the case when some decision variables are continuous.

- Question to be answered. In our example the goal is to find the single best system, where *best* is some criterion of goodness determined by the decision-maker. Other possibilities include finding all systems satisfying a set of criteria or ranking the best r of the k systems.

- Number of performance measures. In our example the comparison of systems is based on only one criterion so the definition of *best* is unambiguous. When multiple performance parameters are considered, the definition of best becomes more subjective. Scatter plots and other graphical aids are particularly useful for comparisons in two, and sometimes more, dimensions.

- Type of performance measures. In our example the single performance measure is estimated by a sample mean of independent observations. Comparisons based on dependent observations are sometimes more efficient, but complicate the statistical analysis. The statistical analysis also would change if the performance measure were estimated with a non-mean, such as a sample standard deviation or quantile.

- Computational cost. In our example the stochastic elements of the system and the system complexity produce a high computational cost, measured typically in elapsed time, money, or both. If the computational cost is not high, then the

comparison is easy—simply simulate each system until the sampling error is negligible.

- Constraints. In our example we make various assumptions about constraints placed on the analysis caused by computing time, analyst time, and analyst sophistication.

- Simulation environment. The combination of computer hardware and software can support or hinder simultaneous runs of various systems, stopping and restarting of the simulations, interactive analysis, graphical analysis, and statistical analysis.

- Common random numbers. Often the k systems have similar logical structure, in which case the use of common random numbers can reduce the computational cost. In our example common random numbers are *not* used; therefore, the data from each system are independent from the other systems' data.

- Use of the results. The choice of analysis method can depend upon whether its purpose is to convince the analyst or another decision-maker.

1.2 The Airline-Reservation Example

We consider $k = 4$ different airline-reservation systems. The single measure of performance is the expected time to failure, E[TTF]—the larger the better. The system works if either of two computers works. Computer failures are rare, repair times are fast, and the resulting E[TTF] is large. The four systems arise from variations in parameters affecting the time-to-failure and time-to-repair distributions. We know from experience that the E[TTF]'s are roughly 100,000 minutes (about 70 days) for all four systems. We are indifferent to expected differences of less than 3000 minutes (about two days).

The large E[TTF]'s, the highly variable nature of rare failures, the similarity of the systems, and (as it turns out) the small *indifference zone* of 3000 minutes yield a problem context with reasonably large computational costs. Although the similarity of the systems suggests the use of variance reduction techniques such as common random numbers, for tutorial simplicity we have agreed to restrict ourselves to independent replications of the systems. In all cases the point estimator for system i is the sample average over the replications allocated to system i by the experiment.

2 INTERACTIVE ANALYSIS

The interactive analysis described here is an *estimation* approach. It considers the $k = 4$ point estimators for the respective E[TTF]'s and the estimates of their sampling errors. The goal is a vague, but well-founded, sense of confidence in the selected system. IA is similar in spirit to the procedure in Schmidt and Taylor (1970, pp. 524–528), except that we suppress their explicit confidence-interval logic. Neither the assumptions nor the cleanly-stated statistical conclusions of R&S and MCPs are found here.

2.1 The Method

The method is to successively experiment with the systems, roughly comparing the four point estimators and their associated standard errors with each other until we are comfortable that the chosen system is the best or negligibly close to the best. A more-detailed discussion of the underlying concepts can be found in Schmeiser (1990).

An interactive driver program reads a specified system i, a set of random-number seeds, a number of microreplications m, and a number of macroreplications b. It runs the simulation model for $n = bm$ replications, producing the TTF observation $Y_{ij\ell}$ on microreplication ℓ of macroreplication j. The driver also produces the point estimator for μ_i, the E[TTF] of system i,

$$\bar{Y}_i = \frac{1}{b}\sum_{j=1}^{b}\bar{Y}_{ij} = \frac{1}{n}\sum_{j=1}^{b}\sum_{\ell=1}^{m}Y_{ij\ell}$$

with associated sample variance of the *macroreplication* estimators

$$S_i^2 = \sum_{j=1}^{b}(\bar{Y}_{ij} - \bar{Y}_i)^2/(b-1)$$

and standard error $se_i = S_i/\sqrt{b}$. (The SERVO software described in Schmeiser and Scott 1991 could be used to automatically compute these statistics.) The relevant part of the output report is

```
..parameters...
..   system                 =  1
..   macroreplications      =  5
..   microreplications      =  5
..monte carlo estimates...
..   estimated E[TTF]       =   111086.
..     with standard error  =    21211.0
```

In this example, twenty-five replications of system 1 have been run. The interpretation is that system 1

has estimated E[TTF] of $11. \times 10^4$ minutes. The last digit included is in the most-significant position of the standard error. The last digit is meaningful only in that $11. \times 10^4$ is a better guess than $1. \times 10^5$. The digits not included are meaningless, since they are dominated by sampling error.

2.2 The Assumptions

The *significant digits* interpretation of the last paragraph is loosely based on the mathematics of the confidence interval on the E[TTF] of system i,

$$\bar{Y}_i \pm t_{b-1,1-\alpha/2} \cdot se_i,$$

which under assumptions of normality and independence of the $\bar{Y}_{ij}$'s covers the true mean μ_i with confidence $1-\alpha$. If these assumptions hold, then for $b > 4$ [$b > 15$] and the typically-used 90% to 99% confidences, the t-values range from 1.6 to 4.6 [1.6 to 2.9]. Since the range of values is small, and since most people choose α arbitrarily, there is little precision lost by stating significant digits rather than a confidence interval. Put another way, little practical (compared to statistical) confidence results if varying α (over the typical values) affects our decision whether to stop or simulate further.

Nevertheless, some care must be taken, since in extreme cases the significant-digit guideline is invalid. Small values of b and α can lead to large t-values; for example, $b = 2$ with a 99% confidence level yields a t-value of 64. In simulations with serially dependent data, large values of b (with fixed n) result in high correlation of the $\bar{Y}_{ij}$'s, typically biasing estimators of the standard error to the low side. Choosing $10 < b < 30$ is often wise (Schmeiser 1982). The independence in our example implies that we can safely choose b relatively large, since only normality of the $\bar{Y}_{ij}$'s (via the central limit theorem) is gained with low values of b.

Interactive experimentation leads to sequential sampling, another source of error in confidence-interval computations. But if the value of b is kept reasonably large, the secondary effects of sequential sampling are negligible in a method that does not specify a confidence level.

2.3 The Example

In this section we summarize a log kept during an analysis of the airline-reservation-system example. The analysis spanned two days.

The initial run, which produces the example output shown above for system 1, is designed just to gain a sense of the magnitude of the required production experiment in terms of time per replication and number of replications. The twenty-five replications require about thirteen minutes on a dedicated SPARCstation 1, so each replication costs about 30 seconds of real time. We extrapolate this time estimate to the other systems, since all four systems are similar.

The standard error estimate of 21000 minutes has only $b - 1 = 4$ degrees of freedom, but that is fine for the following rough analysis. We are interested only in detecting differences between the best system and inferior systems when the differences are in fact at least 3000 minutes. So the standard errors will need to drop to at most 1500 minutes, or 1/14 the current value. Therefore, the number of replications needs to be about 14^2 times longer: 5000 replications. This means that the "worst-case" projected experiment time is 5000 replications times half a minute per replication times four systems. Seven days. The standard error has few degrees of freedom, so the actual experiment time may be from five to nine days.

Rather than begin a seven-day experiment so soon after starting the comparison, and since it is time to go home for the day, we first run an overnight experiment: For each of the four systems, we run $b = 15$ macroreplications of $m = 15$ microreplications; as agreed upon for this tutorial, we use different random numbers for each system.

The overnight results are given in Table 1. Stated in significant digits, $\bar{Y}_1 = 11.1 \times 10^4$, $\bar{Y}_2 = 10.3 \times 10^4$, $\bar{Y}_3 = 9.4 \times 10^4$, and $\bar{Y}_4 = 0.8 \times 10^4$. System 4 appears to be an order of magnitude from being competitive; however, inspection shows an input error, so ignore the system 4 results for now, while being pleased this was an overnight run rather than a week-long run. System 1 looks better than systems 2 and 3, regardless of the indifference value. System 3 already is close to being eliminated, so the total experimentation is looking shorter than the worst-case projections.

Table 1: IA Overnight-Experiment Results

i	1	2	3	4
$\bar{Y}_i$	110762.	103265.	93968.	7994.
se_i	5757.	4622.	5226.	538.

Now would be a good time to create some graphics, e.g., boxplots of the macroreplication estimators $\bar{Y}_{ij}$. (We have not included graphics in this paper due to space constraints.)

Needing to go home in about an hour, let us devote about 30 minutes exclusively to system 4, for which we take $b = 10$ and $m = 6$. It turns out that $\bar{Y}_4 =$

8.8×10^4 minutes, which is not competitive, especially if we remember the indifference value of 0.3×10^4 minutes.

We will be back in about five hours, so we have time to make a run of systems 1 and 2, each for $b = 30$ and $m = 10$. The results are $\bar{Y}_1 = 10.9 \times 10^4$ and $\bar{Y}_2 = 10.5 \times 10^4$. This is more cumulative evidence indicating system 1 is best, but the evidence from this run alone is not conclusive. Since the additional information lowers the standard error of the estimators for systems 1 and 2, we decide that it is safe to eliminate system 3 from further consideration. In fact, since the indifference value is 3000 minutes, we could at this point select system 1 as best, albeit with low (undefined, subjective) confidence.

We are now far ahead of schedule, which (arbitrarily for this tutorial) is to finish before the Memorial Day weekend. With the extra time, we devote another night's run to the comparison. We simulate systems 1, 2, and 4, each for $b = 30$ and $m = 20$. (System 4 is included since the time is available; our initial look was probably sufficient, despite it being only sixty microreplications.) The results are $\bar{Y}_1 = 10.6 \times 10^4$, $\bar{Y}_2 = 10.4 \times 10^4$, and $\bar{Y}_4 = 9.1 \times 10^4$. System 4 is clearly noncompetitive. Our confidence in system 1 has risen yet more, since it is again best, although still not conclusively. With the extra assurance of the 3000-minute indifference zone, we declare system 1 our choice with a confidence easily great enough for this tutorial application.

2.4 Discussion

IA carries a variety of advantages compared to the formal methods such as those to be discussed in the next sections. Noncompetitive systems are eliminated quickly in IA, and so require less computation than better systems; many formal methods assume the worst case and devote equal numbers of observations to each system. IA allows heavy computation to be conveniently scheduled while the analyst is away; many formal methods require an inconvenient single large run. IA forces the analyst to think about the results, which helps to detect mistakes. IA extends directly to estimators other than means (Schmeiser, Avramidis, and Hashem 1990); new mathematics is required for most formal methods. Common random numbers can be incorporated into IA by computing standard errors on differences or by using the same analysis with some increased confidence in the results. Adding new systems is easy with IA's informality; most formal methods require the number of systems k to be fixed.

Three disadvantages exist. First, the lack of a precise confidence statement causes discomfort for many people. Second, the analyst ignoring or misusing the standard errors allows incorrect conclusions. Third, using significant digits as a measure of sampling error is crude, since a significant digit is added only by simulating 100 times longer. Of course, the standard errors are available for more precise computation if the analyst desires.

3 RANKING AND SELECTION

Ranking and selection procedures are statistical methods specifically developed to select the best system from a set of competing systems. Provided certain assumptions are met, these methods usually guarantee that the probability of a correct selection will be at least some user-specified value. This section discusses the normal means procedure of Rinott (1978). We then apply the Rinott procedure to the airline-reservation-system problem at hand.

3.1 The Method

The general goal behind R&S methods is to select the best system from among k competitors. Here, we have $k = 4$ airline-reservation systems. By *best*, we mean the system having the largest underlying E[TTF]. Suppose we denote the E[TTF] arising from system i by μ_i, $i = 1, 2, \ldots, k$, and the associated ordered μ_i's by $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]}$. (We assume that the μ_i's, $\mu_{[i]}$'s, and their pairings are completely unknown.) Since we prefer the E[TTF] to be as large as possible, the mean difference between the two best systems in the ongoing airline-reservation example is $\mu_{[k]} - \mu_{[k-1]}$. The smaller this difference is, the greater the amount of sampling that will be required to differentiate between the two best systems. Of course, if $\mu_{[k]} - \mu_{[k-1]}$ is very small, say less than $\delta > 0$, then for all practical purposes, it would not matter which of the two associated systems we chose as best. In other words, we regard δ as the smallest difference "worth detecting." In the airline-reservation example, we have taken $\delta = 3000$ minutes.

We would also like to be assured that the probability that we make a *correct selection* (CS) of the best system will be at least a certain high value, P^*. The greater the value of P^*, the greater the number of observations that will be required. We take $P^* = 0.90$ in our example.

We will apply a two-stage procedure due to Rinott (1978) to the airline-reservation problem. The procedure assumes that system i produces independent and identically distributed (i.i.d.) normal (μ_i, σ_i^2) output, where μ_i and σ_i^2 are unknown, $i = 1, 2, \ldots, k$,

and where the k systems are independent. If $\mu_{[k]} - \mu_{[k-1]} > \delta$, the procedure guarantees that $P\{\text{CS}\} \geq P^*$.

The procedure runs as follows. In the first stage of sampling, we take a random sample of b_0 observations from each of the k normal populations. We use as our observations from system i the *macroreplication* estimators, $\bar{Y}_{i1}, \bar{Y}_{i2}, \ldots$, defined as in Subsection 2.1; for now, we will assume that they are i.i.d. normal. Calculate the first-stage sample means,

$$\bar{Y}_i^{(1)} = \sum_{j=1}^{b_0} \bar{Y}_{ij}/b_0,$$

and sample variances

$$S_i^2 = \sum_{j=1}^{b_0} (\bar{Y}_{ij} - \bar{Y}_i^{(1)})^2/(b_0 - 1),$$

for $i = 1, 2, \ldots, k$. The sample variances are used to determine the number of observations (macroreplications) which must be taken in the second stage of sampling; the larger a sample variance, the more macroreplications must be taken in the second stage from the associated system. Now set $b_i = \max\{b_0, \lceil (hS_i/\delta)^2 \rceil\}$, where $\lceil \cdot \rceil$ is the "ceiling" function, and h is a constant that solves a certain integral, and is tabled in, e.g., Wilcox (1984). During the second stage of sampling, take $b_i - b_0$ *additional* observations from the ith system, $i = 1, 2, \ldots, k$.

Finally, we calculate the grand means $\bar{Y}_i = \sum_{j=1}^{b_i} \bar{Y}_{ij}/b_i$, $i = 1, 2, \ldots, k$, and select the system having the largest $\bar{Y}_i$ as best (which is certainly intuitively appealing).

3.2 The Assumptions

Rinott's procedure requires that the observations (macroreplications) taken within a particular system be i.i.d. normal. We discuss these assumptions in this subsection.

The macroreplication estimators, $\bar{Y}_{i1}, \bar{Y}_{i2}, \ldots, \bar{Y}_{ib_i}$, from the ith system are assumed to be i.i.d. with expectation μ_i. This is trivially true since the macroreplications are independent of each other.

The macroreplications from across all systems are assumed to be independent; i.e., if $i \neq i'$, all $\bar{Y}_{ij}$'s are independent of all $\bar{Y}_{i'j}$'s, $j = 1, 2, \ldots$. This requirement is also satisfied trivially since different random-number streams are chosen for each system's simulation. (See Subsection 4.5 for a discussion concerning the use of common random numbers.)

The macroreplication estimators, $\bar{Y}_{ij}$, for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, b_i$, are assumed to be normally distributed. If the number of microreplications m is large enough, say at least 20, then the central limit theorem yields approximate normality for the macroreplication estimators.

We make no assumptions on the variances of the macroreplications. Although there are a number of R&S procedures devised for the special case of normal populations with unknown but common variance, we will not resort to those procedures here.

3.3 The Example

Our goal is to find the system having the largest E[TTF]. To achieve the goal the following sequence of experiments was performed:

1. A debugging experiment to check the computer code and assess execution speed.
2. A pilot experiment to study characteristics of the data and aid in planning the production run.
3. A production run to produce the final results.

All R&S experiments and analyses were performed on various SPARCstations.

3.3.1 Debugging Experiment

Five macroreplications, each consisting of five microreplications of system 1, produced a sample mean TTF of 129182. minutes and a sample standard deviation 69417.2 minutes. Each microreplication took about 24 seconds of real time on a (non-dedicated) SPARCstation 1. Since the sample variance was so large, we decided to conduct a somewhat larger pilot study; this would also serve as the first stage of the Rinott procedure. The pilot study would take 20 macroreplications, each consisting of 20 microreplications, for each of the $k = 4$ systems. We anticipated that the pilot study would use at most 10 hours of real time.

3.3.2 Pilot Experiment

By dividing the pilot study among various SPARCstations, we were able to complete it in less than 3 hours. The results are given in Table 2.

To check our normality assumption, we used the pilot study to conduct Shapiro-Wilks tests on the 20 macroreplications from each system; the tests passed at all reasonable levels. We mention in passing that we could have also conducted Bartlett's test to check for equality of variances among the systems; we did not do so since Rinott's procedure has no restrictions on the variances of the macroreplication estimators. We also remark that an IA interpretation of the pilot

Table 2: R&S Pilot Experiment ($b_0 = 20$)

i	1	2	3	4
$\bar{Y}_i^{(1)}$	108286.	107686.	96167.7	89747.9
S_i	29157.3	24289.9	25319.5	20810.8
se_i	6519.8	5431.4	5661.6	4653.4
b_i	699	485	527	356

study would have immediately eliminated system 4 (and maybe even system 3) from further consideration.

For the case $k = 4$ and $P^\star = 0.90$, the critical constant from Wilcox (1984) is $h = 2.720$. This enabled us to calculate the b_i-values for the second stage of sampling. Since the pilot study was also intended to be used as the first stage of Rinott sampling, we have displayed the resulting b_i-values in the table. For example, for system 2, we needed to take $b_2 - b_0 = 465$ additional macroreplications in stage two, each consisting of $m = 20$ microreplications. The total number of microreplications over all four systems was about 40000. A worst-case scenario of 24 seconds of real time per microreplication (as in the debugging experiment) implied that the production run might take 250 hours; luckily, we had access to a number of SPARCstations, some faster than the SPARCstation 1.

3.3.3 Final Results

By dividing the production runs among the various SPARCstations, we were able to complete them in less than 2 days. The results are given in Table 3. These results clearly establish system 1 as the winner. We can make the formal statement that we are at least 90% sure that we have made the correct selection (with the proviso that the true difference between the best and second best E[TTF]'s is at least $\delta = 3000$ minutes).

Table 3: R&S Production Run

i	1	2	3	4
$\bar{Y}_i$	110816.5	106411.8	99093.1	86568.9
se_i	872.0	1046.5	894.2	985.8

3.4 Discussion

There are a number of reasons to use R&S techniques when seeking the best of a number of competing systems. Procedures such as Rinott's guarantee the user of a correct selection with high probability when the true difference between the best and second best system is at least δ; even when the true difference is less than δ, Rinott's procedure insures selection with high probability of a "good" system (i.e., one that is within δ of the best). This guarantee compares favorably to the simple "yes" or "no" answer that a classical hypothesis test is likely to provide. R&S procedures are also easy to use, as our Rinott example demonstrated; little more than one tabled constant look-up and a sample-mean calculation is required.

One drawback to Rinott's procedure is that it tends to be conservative, i.e., it sometimes takes more observations than necessary in the presence of "favorable" system mean configurations (i.e., configurations in which the largest mean and the others differ by more than δ). This drawback arises from the fact that Rinott guarantees $P\{\text{CS}\} \geq P^\star$ for *all* configurations of the system means for which the best is at least δ better than the second best. But Rinott is just one of many R&S techniques for the normal means problem. For instance, if we were to assume (or force) common variances among the competing systems, we could use a variety of more efficient R&S procedures, some of which enjoy the capability of sequentially eliminating systems deemed as noncompetitive; these procedures capitalize on favorable system mean configurations by terminating sampling early. Bechhofer, Dunnett, Goldsman, and Hartmann (1990) study a number of such procedures. For additional reading, Gibbons, Olkin, and Sobel (1977) is a nice introductory R&S text, and Law and Kelton (1991) shows how to apply R&S techniques in a simulation context.

4 MULTIPLE COMPARISONS

Multiple-comparison procedures treat the optimization problem as an inference problem on the performance parameters of interest. MCPs account for the error that arises when making simultaneous inferences about differences in performance among k systems. This section introduces a specific MCP, multiple comparisons with the best (MCB), and applies it to the airline-reservation-system example.

4.1 The Method

MCB provides simultaneous statistical inference on either $\mu_i - \max_{\ell \neq i} \mu_\ell$ or $\mu_i - \min_{\ell \neq i} \mu_\ell$, for $i = 1, 2, \ldots, k$, where μ_i is the performance parameter of

system i. In the airline-reservation example, μ_i is the E[TTF] for computer configuration i.

This collection of parameters is particularly relevant for optimization. For example, suppose a larger performance parameter is better (as in the airline-reservation system). If $\mu_i - \max_{\ell \neq i} \mu_\ell > 0$, then system i is the best, because all other systems have smaller performance parameter. On the other hand, if $\mu_i - \max_{\ell \neq i} \mu_\ell < 0$, then system i is not the best, since there is another system with larger performance parameter. Even when $\mu_i - \max_{\ell \neq i} \mu_\ell < 0$, if $\mu_i - \max_{\ell \neq i} \mu_\ell > -\delta$, where δ is a positive number, then system i is within δ of the best.

If optimization is the goal, then inference on the k parameters $\mu_i - \max_{\ell \neq i} \mu_\ell$, $i = 1, 2, \ldots, k$, is superior to inference on the $k(k-1)/2$ parameters $\mu_i - \mu_\ell$, $\forall i \neq \ell$. Simultaneous inference is typically sharper (smaller differences in performance can be detected) when fewer parameters are included in the inference.

The parametric version of MCB described here assumes that the output data from the simulation experiment is approximated by the oneway analysis-of-variance model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

for systems $i = 1, 2, \ldots, k$ and replications $j = 1, 2, \ldots, n_i$, where the ε_{ij}'s are i.i.d. normal $(0, \sigma^2)$ random variables. (We make no distinction between micro and macroreplications for the moment.)

Under model (1), Hsu (1984) derived simultaneous $(1-\alpha)100\%$ confidence intervals for $\mu_i - \max_{\ell \neq i} \mu_\ell$ for $i = 1, 2, \ldots, k$. In the "balanced" case ($n_1 = n_2 = \cdots = n_r = n$), the form of the ith interval is

$$\left(\bar{Y}_i - \max_{\ell \neq i} \bar{Y}_\ell - dS\sqrt{\frac{2}{n}}\right)^- , \left(\bar{Y}_i - \max_{\ell \neq i} \bar{Y}_\ell + dS\sqrt{\frac{2}{n}}\right)^+$$

where $\bar{Y}_i$ is the sample mean of the outputs from system i, S^2 is a pooled estimator of σ^2, $d = d_{1-\alpha,k(n-1),k}$ is a critical value, $x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$. The intervals for the "unbalanced" case are similar, but difficult to write compactly.

Notice that the MCB intervals are constrained intervals, meaning that each interval either contains 0 or one of its endpoints is 0. In a maximization problem, if the confidence interval for $\mu_i - \max_{\ell \neq i} \mu_\ell$ contains 0 it means that—relative to the sampling error in the point estimators—system i is not significantly different from the best system, and may be the best. If the upper endpoint of the interval is 0, then system i is not the best system. On the other hand, if the lower endpoint is 0, then system i is the best system; at most one system will have lower endpoint 0. These statements are made with confidence level $1 - \alpha$.

Hochberg and Tamhane (1987) describe the theoretical development and practical application of MCB. Nelson (1992) gives a detailed algorithm for the balanced case. Both references provide tables of critical values. Statistical analysis packages that compute MCB intervals include JMP (version 2) and Minitab (release 8).

4.2 The Assumptions

The assumptions implied by model (1), and their interpretation in simulation experiments, are discussed in this subsection.

The output data from within each system i, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, are assumed to be i.i.d. with common expectation μ_i. This will be true if the outputs are from replications (as in the airline-reservation example), or approximately true if they are batch means from within a single replication of a stationary process.

The output data from across all systems on replication j, $Y_{1j}, Y_{2j}, \ldots, Y_{kj}$, are assumed to be independent. This will be true if different random number streams or seeds are chosen for the simulation of each system. Since simulators sometimes use common random numbers (CRN) to sharpen comparisons, we comment on the effect of CRN on MCB in Subsection 4.5.

All the outputs Y_{ij}, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$, are assumed to be normally distributed. This may be approximately true if each Y_{ij} is an average of many outputs.

All the outputs Y_{ij}, for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$, are assumed to have common variance $\sigma^2 = \text{Var}[Y_{ij}]$, $\forall i, j$. There is no reason to believe this assumption holds in general.

The assumptions of normally distributed data and common variance are clearly the most tenuous. However, since simulators can often obtain a large number of replications, n, the method of batch means can be used to improve the approximation to both assumptions.

Consider the outputs from system i in the balanced case: $Y_{i1}, Y_{i2}, \ldots, Y_{in}$. If $b_i m_i = n$, then we can transform the data into b_i batch means of m_i outputs as follows:

$$\bar{Y}_{i\ell} = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{i,(\ell-1)m_i+j} \tag{2}$$

for $\ell = 1, 2, \ldots, b_i$. Notice that the "batch means" are identical to the macroreplication estimators described in the previous sections. We use different terminology because we "batch" the data *after* collecting it to improve the approximations of normality and equal variance.

Let $\sigma_i^2 = \text{Var}[Y_{ij}]$. Then $\text{Var}[\bar{Y}_{i\ell}] = \sigma_i^2/m_i$; that is, the variance of the batch means can be controlled by the choice of batch size, m_i. In addition, the batch means will tend to be normally distributed due to the central limit theorem effect. Therefore, by choosing the batch sizes $m_1, m_2, \ldots, m_k$ (equivalently the numbers of batches $b_1, b_2, \ldots, b_k$) appropriately, the batch means $\bar{Y}_{i\ell}$ will be more nearly normally distributed, and the variances of the batch means more nearly equal, than the original data Y_{ij}.

A drawback of batching is the loss of degrees of freedom—from $\sum_{i=1}^{k}(n_i - 1)$ to $\sum_{i=1}^{k}(b_i - 1)$—which affects how sharp the inference is. Goldsman and Nelson (1990) analyzed the effect of batching on MCB and found that, for $3 \leq k \leq 10$ systems, very little is lost as long as the number of batches is $b_i \geq 20$; this result holds no matter how many replications, n, are available. Therefore, batching can (and should) be used provided the simulator has enough data to form at least 20 batch means.

4.3 Planning

The IA approach Schmeiser advocates proceeds until the estimation error is small enough that a winner can be declared. The R&S procedure Goldsman advocates is also designed to yield a winner. MCPs provide information about the relative performance of the systems, but they are typically not designed to produce a winner. However, careful planning of the experiment can insure that useful results are obtained.

Suppose that a difference in performance of more than δ is considered important. Hsu (1988) derived an expression for the sample size, n, required to guarantee that, with high probability, the system that MCB infers to be the best is in fact within δ of the true best. This "power" calculation requires an estimate of δ/σ, and is similar in spirit to the second-stage sample-size calculation of Rinott's procedure.

Since software for Hsu's sample-size calculation is not readily available, a crude approximation is to take

$$n \geq \left(\frac{\sqrt{2} d_{1-\alpha,k(n-1),k}\, \sigma}{\delta} \right)^2 \qquad (3)$$

which approximates the sample size required to obtain a confidence-interval halfwidth less than or equal to δ. An estimate of σ may be obtained from a pilot study, and a conservative (small) value of n can be chosen to determine a value of $d_{1-\alpha,k(n-1),k}$ to insert in the formula.

4.4 The Example

The outline of experiments we performed is the same as that described in Subsection 3.3. The data was written to a file, and the analysis was conducted using Splus. All experiments and analyses were performed on a DECstation 3100.

4.4.1 Debugging Experiment

Ten replications of system 1 produced a sample mean TTF of $\bar{Y}_1 = 60761.$ minutes and a sample standard deviation of $S = 71450.$ minutes. Each replication took about 3 seconds of real time. Under the (arbitrary) constraint of running and analyzing the pilot experiment within an hour, this implied that a pilot experiment of $n = 200$ replications for each system could be performed (approximately 40 minutes of real time to generate the data).

4.4.2 Pilot Experiment

For each system, $n = 200$ replications of TTF were generated. Different random number seeds were used to initialize the experiment for each system to obtain independent data, as assumed by model (1). The simulations actually used nearly the entire hour.

Histograms and quantile-quantile plots of the data from each system indicated that the data was not normally distributed. Boxplots of the data indicated that system 1 was somewhat more variable than the others; the ratio of the largest sample variance (system 1) to the smallest sample variance (system 4) was almost 2.5.

To improve the approximation of normality, the data from each system were batched into $b = 40$ batch means of $m = 5$ outputs. Visually the data appeared to be more normally distributed after the transformation. Using the batch means, 95% simultaneous MCB confidence intervals were formed for $\mu_i - \max_{\ell \neq i} \mu_\ell$, $i = 1, 2, 3, 4$; the results are displayed in Table 4.

Table 4: MCB Pilot Experiment

i	lower limit	$\bar{Y}_i - \max_{\ell \neq i} \bar{Y}_\ell$	upper limit
1	-11349.	7091.	25530.
2	-31242.	-12802.	5638.
3	-25530.	-7091.	11349.
4	-36172.	-17732.	708.

System 1 was the sample best, but since all the intervals contained 0 no system could be declared

to be the best. The lower limit on the interval for $\mu_1 - \max_{\ell \neq 1} \mu_\ell$ indicated that, if system 1 is not the best, its E[TTF] could be as much as 11349. minutes less than the best; since this was greater than the indifference zone of $\delta = 3000$ minutes, a "production run" was planned.

Using the pooled standard deviation estimate from the pilot run, $S = 92449.$, the indifference zone $\delta = 3000$, and the critical value $d_{0.95,4(40),4} = 2.078$, formula (3) predicted that $n \approx 8200$ replications would be required to distinguish differences of 3000 minutes. Approximately 42 hours of real time would be required for the computer to complete that many replications.

Rather than save all of the TTF data from the production run, the outputs were batched as they were generated into $b = 300$ batch means of $m = 27$ outputs each (implying $n = 8100$ replications; the 200 pilot replications were available to add to the data set if needed). Having 300 batch means allowed some flexibility for further batching to achieve approximately equal variances while still keeping the number of batches above 20. The production runs were executed over a weekend.

4.4.3 Final Results

The batch means from all four systems appeared to be normally distributed. Applying Bartlett's test for equality of variance yielded a test statistic of 13.11, which is larger than the 0.995 quantile of a χ^2 random variable with 3 degrees of freedom; the variances did not appear to be equal. However, the test statistic without system 4 was only 0.99, which is not significant.

The ratio of the pooled sample variance from systems 1, 2, and 3 to the sample variance of system 4 was about 1.5. Therefore, approximately equal variances could be obtained by rebatching the 300 batch means from systems 1, 2, 3 and 4 into $b_1 = b_2 = b_3 = 100$ and $b_4 = 150$ batch means, respectively. Bartlett's test on the rebatched data yielded a test statistic of 1.03, which is not significant.

The 95% simultaneous MCB confidence intervals based on the rebatched data are given in Table 5. System 1 was conclusively identified as the best system (with confidence level 0.95), and the best guess is that it is superior by 5288. minutes. System 2 was the sample second best, but it may be inferior to system 1 by as much as 8284. minutes, which is greater than the indifference zone of 3000 minutes.

Table 5: MCB for Production Run

i	lower limit	$\bar{Y}_i - \max_{\ell \neq i} \bar{Y}_\ell$	upper limit
1	0	5288.	8284.
2	-8284.	-5288.	0
3	-12747.	-9751.	0
4	-25409.	-22675.	0

4.5 Discussion

MCPs recognize that selecting the best system is a multivariate-estimation problem, and they explicitly account for the joint (overall) error inherent in making statements about multiple performance parameters. Informal methods for measuring sampling error do not account for the possibility of simultaneous errors in different directions. MCPs provide inference about not only the best system, but also relationships among all the systems. In the case of MCB, the difference between the expected performance of each system and the best of the other systems is bounded. These insights are useful when the performance parameter of interest (E[TTF] in the example) does not account for all of the differences among the systems (e.g., cost of installation). MCPs can provide inference from a single stage of sampling, although, as the example illustrates, experiment planning may be required to guarantee useful results since a winner is not guaranteed.

The primary disadvantage of MCPs is their reliance on rather strict distributional assumptions. Nonparametric procedures exist that remove assumptions (such as normality) on the marginal distributions, but they typically have lower power to discriminate differences. A particularly nettlesome limitation is the assumption of independence across systems, which rules out the use of CRN. CRN is a variance reduction technique for sharpening estimators of differences. The improvement is obtained by inducing positive dependence across replications from each system, $Y_{1j}, Y_{2j}, \ldots, Y_{kj}$, which violates an assumption of model (1).

We have been able to show that MCB is conservative, under fairly general conditions, when CRN is employed, which means that a true confidence level greater than $1 - \alpha$ is achieved. Ideally, the MCP should incorporate CRN to make the inference sharper while preserving the desired confidence level. Some progress has been made in developing MCPs that incorporate CRN (e.g., Yang and Nelson 1991).

ACKNOWLEDGMENTS

David Goldsman's work was supported by National Science Foundation Grant No. DDM-9012020. Barry Nelson's work was supported by National Science Foundation Grant No. DDM-8922721. Bruce Schmeiser's work was supported by National Science Foundation Grant No. DMS-8717799. The authors acknowledge the helpful comments of Jon R. Hill, Jason C. Hsu, Yu-Hui Tao, and Jin Wang.

REFERENCES

Bechhofer, R. E., C. Dunnett, D. Goldsman, and 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. *Communications in Statistics—Simulation and Computation* **B19**, 971–1006.

Gibbons, J. D., I. Olkin, and M. Sobel. 1977. *Selecting and Ordering Populations: A New Statistical Methodology.* New York: John Wiley.

Goldsman, L., and B. L. Nelson. 1990. Batch-size effects on simulation optimization using multiple comparisons with the best. In: *Proceedings of the 1990 Winter Simulation Conference*, eds. O. Balci, R.P. Sadowski, and R.E. Nance, 288–293. Institute of Electrical and Electronics Engineers.

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

Hsu, J. C. 1984. Ranking and selection and multiple comparisons with the best. In: *Design of Experiments: Ranking and Selection*, eds. T. J. Santner and A. C. Tamhane, 23–33. New York: Marcel Dekker.

Hsu, J. C. 1988. Sample size computation for designing multiple comparison experiments. *Computational Statistics & Data Analysis* **7**, 79–91.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis.* New York: McGraw-Hill.

Nelson, B. L. 1992. Statistical analysis of simulation results. In: *Handbook of Industrial Engineering*, Second Edition, ed. G. Salvendy, in press. New York: John Wiley.

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

Schmeiser, B. W. 1982. Batch size effects in the analysis of simulation output. *Operations Research* **30**, 556–568.

Schmeiser, B. W. 1990. Simulation experiments. In: *Handbook of Operations Research and Management Science, Volume 2: Stochastic Models*, eds. D. Heyman and M. Sobel, 295–330. Amsterdam: North Holland.

Schmeiser, B. W., T. Avramidis, and S. Hashem. 1990. Overlapping batch statistics. In: *Proceedings of the 1990 Winter Simulation Conference*, eds. O. Balci, R. P. Sadowski, and R. E. Nance, 395–398. Institute of Electrical and Electronics Engineers.

Schmeiser, B. W., and M. D. Scott. 1991. SERVO: Simulation experiments with random-vector output. In: *Proceedings of the 1991 Winter Simulation Conference*, eds. B. L. Nelson, G. M. Clark, and W. D. Kelton, this volume. Institute of Electrical and Electronics Engineers.

Schmidt, J. W., and R. E. Taylor. 1970. *Simulation and Analysis of Industrial Systems.* Homewood, Illinois: Richard D. Irwin.

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

Yang, W., and B. L. Nelson. 1991. Using common random numbers and control variates in multiple-comparison procedures. *Operations Research* **39**, in press.

AUTHOR BIOGRAPHIES

DAVID GOLDSMAN is an Associate Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis and ranking and selection. He is Secretary-Treasurer of the TIMS College on Simulation.

BARRY L. NELSON is an Associate Professor in the Department of Industrial and Systems Engineering at The Ohio State University. His research interests are experiment design and analysis of stochastic simulations. He is Vice President of the TIMS College on Simulation, and is *Proceedings* Editor for the 1991 Winter Simulation Conference.

BRUCE SCHMEISER is a Professor in the School of Industrial Engineering at Purdue University. His research interests include input modeling, random-variate generation, output analysis, and variance reduction. He is the current Simulation Area Editor of *Operations Research* and a Member of the Council of the Operations Research Society of America. He is an active participant in the Winter Simulation Conference, including being Program Chairman in 1983 and Chairman of the Board of Directors during 1988-1990.