Sequential Selection Procedures: Using Sample Means to Improve Efficiency

E. Jack Chen\textsuperscript{a}, W. David Kelton\textsuperscript{b}

\textsuperscript{a} BASF Corporation, Mount Olive, NJ 07828-1234, USA
Tel: 973-426-6047, Email: chenej@basf.com

\textsuperscript{b} Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, OH 45221-0130, USA
Tel: 513-556-6834, Email: david.kelton@uc.edu

Received 16 May 2002, revised 12 March 2003

Two-stage selection procedures have been widely studied and applied to determine appropriate sample sizes for selecting the best of \( k \) designs. However, standard “indifference-zone” procedures are derived with a statistically conservative least-favorable-configuration assumption. The Enhanced Two-Stage Selection (ETSS) is a procedure that takes into account not only the variance of samples but also the difference between sample means when determining the sample sizes. This paper discusses an implementation of sequential ranking and selection procedures due to the ETSS procedure to avoid relying too much on information obtained in just one stage. We show that the needed ratios of sample sizes to maximize the probability of correct selection is approximately maintained at all iterations. An experimental performance evaluation demonstrates the efficiency of our sequential procedures.

(Simulation; Ranking and Selection; Sample-Size Allocation; Indifference Zone)

1. INTRODUCTION

Discrete-event simulation has been widely used to compare alternative system designs or operating policies. When evaluating \( k \) alternative system designs, we would like to select one as the best (or the best \( b \) designs) and to control the probability that the selected design really is the best, or to maximize the probability of correct selection with a pre-determined computing budget. Let \( \mu_i \) denote the expected response of design \( i \). Our goal is to find a design with the smallest expected response \( \mu^* = \min_{1 \leq i \leq k} \mu_i \). If the goal is to select a design with the biggest expected response, just replace \( \min \) with \( \max \) in the above. We achieve this
goal by using a class of ranking and selection (R&S) procedures. However, efficiency is still a key concern for using simulation to solve R&S problems.

Many R&S procedures are directly or indirectly based on Dudewicz and Dalal’s (1975) or Rinott’s (1978) indifference-zone-selection procedures. However, these indifference-zone-selection procedures determine the number of additional replications based on a conservative least-favorable-configuration (LFC) assumption and do not take into account the value of sample means; see Section 2.2. Consequently, those procedures are suitable only when the number of designs under consideration is small. It is well known that a subset pre-selection can improve the efficiency of two-stage selection procedures; see Section 2.4. However, the same drawback exists when we apply those two-stage selection procedures on designs surviving the pre-selection.

Some new approaches are OCBA (Optimal Computing Budget Allocation, Chen et al. 2000), ETSS (Chen and Kelton 2000) and adjusted ETSS (Chen 2002) and incorporate first-stage sample mean information in determining the number of additional replications. Both OCBA and ETSS procedures show that an average-case analysis can lead to a significant reduction in computing effort, relative to Rinott’s procedure. There are many other new approaches aimed at improving the efficiency of R&S procedures; Berger and Deely (1994), Gupta and Miescke (1994, 1996), and Chick and Inoue (2001) use a Bayesian framework for constructing R&S procedures. Nelson et al. (2001) develop procedures for selecting the best when the number of alternative is large. Some drawbacks of two-stage selection procedures are that they rely heavily on the information obtained from only the first-stage, and the resources available may not be able to allocate the full sample sizes required for the second stage. Chen et al. (2000), Chick and Inoue (2001), and Kim and Nelson (2001) have developed sequential R&S procedures. For an overview of existing methods of R&S see Law and Kelton (2000).

Let CS denote the event of “correct selection.” In a stochastic simulation, CS can never be guaranteed with certainty. The possibility of CS, denoted by P(CS), depends on sample sizes and other uncontrollable factors. Chen et al. (2000) show that P(CS) is approximately maximized, or the sample size is approximately minimized, if the sample size for each design satisfies certain ratios (see Section 2.4). We propose a simple heuristic sequential R&S procedure such that the allocated sample size for each design approximately satisfies those ratios at each iteration.

The rest of this paper is organized as follows. In Section 2, we provide background necessary for the proposed procedure. In Section 3, we present the methodologies and the proposed
R&S procedure. In Section 4, we give some empirical-experimental results. In Section 5, we give concluding remarks.

2. BACKGROUND

First, some notation:

$X_{ij}$: the observation from the $j^{th}$ replication or batch of the $i^{th}$ design,

$r$: the intermediate number of replications or batches at a particular iteration,

$N_i$: the total number of replications or batches for design $i$,

$\mu_i$: the expected performance measure for design $i$, i.e., $\mu_i = E(X_{ij})$,

$\bar{X}_i(r)$: the sample mean performance measure for design $i$, i.e., $\sum_{j=1}^{r} X_{ij}/r$,

$\bar{X}_i$: the running sample mean of design $i$ at certain iteration (shorthand for $\bar{X}_i(r)$),

$\sigma_i^2$: the variance of the observed performance measure of design $i$ from one replication or batch, i.e., $\sigma_i^2 = \text{Var}(X_{ij})$,

$S_i^2(r)$: the sample variance of design $i$ with $r$ replications or batches, i.e., $S_i^2(r) = \sum_{j=1}^{r} (X_{ij} - \bar{X}_i(r))^2/(r-1)$.

2.1 Indifference-Zone-Selection Procedures

Let $\mu_{i_l}$ be the $l^{th}$ smallest of the $\mu_i$'s, so that $\mu_{i_1} \leq \mu_{i_2} \leq \ldots \leq \mu_{i_k}$. Our goal is to select a design with the smallest expected response $\mu_{i_1}$. In practice however, if $\mu_{i_1}$ and $\mu_{i_2}$ are very close together, we might not care if we mistakenly choose design $i_2$, whose expected response is $\mu_{i_2}$. The “practically significant” difference $d^*$ (a positive real number) between the best and next-best design is called the indifference zone in the statistical literature and represents the smallest difference about which we care. Therefore, we want a procedure that avoids making a large number of replications or batches to resolve differences less than $d^*$. That is, we want $P_{(CS)} \geq P^*$ provided that $\mu_{i_2} - \mu_{i_1} \geq d^*$, where the minimal CS probability $P^*$ and the “indifference” amount $d^*$ are both specified by the user.
2.2 The Two-Stage Rinott Procedure

The two-stage procedure of Rinott (1978) has been widely studied and applied. Let \( n_0 \) be the number of initial replications or batches. The first-stage sample means \( \bar{X}_i(n_0) = \frac{\sum_{j=1}^{n_0} X_{ij}}{n_0} \), and marginal sample variances

\[
S^2_i(n_0) = \frac{\sum_{j=1}^{n_0} (X_{ij} - \bar{X}_i(n_0))^2}{n_0 - 1},
\]

for \( i = 1, 2, \ldots, k \) are computed. Based on the number of initial replications or batches \( n_0 \) and the sample variance \( S^2_i(n_0) \) obtained from the first stage, the number of additional simulation replications or batches for each design in the second stage is \( N_i - n_0 \), where

\[
N_i = \max(n_0, \lceil (hS_i(n_0)/d^*)^2 \rceil), \quad \text{for } i = 1, 2, \ldots, k,
\]

(1)

where \( \lceil z \rceil \) is the smallest integer that is greater than or equal to the real number \( z \), and \( h \) (which depends on \( k, P^*, \) and \( n_0 \)) is a constant that solves Rinott’s (1978) integral (\( h \) can be calculated by the FORTRAN program rinott in Bechhofer et al. (1995), or can be found from the tables in Wilcox (1984)). We then compute the overall sample means \( \bar{X}_i(N_i) = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i} \), and select the design with the smallest \( \bar{X}_i(N_i) \) as the best one. Basically, the computing budget is allocated proportionally to the estimated sample variances. Moreover, the derivation of this procedure is (conservatively) based on the LFC, i.e., assuming \( \mu_{il} = \mu_{i1} + d^* \), for all \( l = 2, \ldots, k \). The LFC is also known as the slippage configuration. However, in reality we rarely encounter the LFC; consequently, this procedure is conservative.

2.3 Optimal Computing Budget Allocation (OCBA)

Chen et al. (2000) propose the OCBA procedure, which is based on a fixed total computing budget \( T = \sum_{i=1}^{k} N_i \) and attempts to maximize \( P(\text{CS}) \). The sequential procedure utilizes the information of both the sample means and sample variances obtained from each iteration to allocate an incremental sample size for each design.

They use the Approximate Probability of Correct Selection (APCS) as a lower bound on the \( P(\text{CS}) \). That is,

\[
P(\text{CS}) \geq 1 - \sum_{i=2}^{k} P[\bar{X}_{i1}(N_{i1}) > \bar{X}_{i2}(N_{i2})].
\]

The right-hand side of the above equation is the APCS. They show that for a fixed number of replications or batches, the APCS can be asymptotically maximized when

\[
\frac{N_i}{N_j} = \left( \frac{\sigma_i/\delta_{i,b}}{\sigma_j/\delta_{j,b}} \right)^2, \quad i, j \in \{1, 2, \ldots, k\}, \quad \text{and } i \neq j \neq b,
\]

(2)
\( N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^{k} \frac{N_i^2}{\sigma_i^2}} \) \hspace{1cm} (3)

where \( \delta_{i,b} = \bar{X}_i(N_i) - \bar{X}_b(N_b) \), \( \bar{X}_b(N_b) = \min_{1 \leq i \leq k} \bar{X}_i(N_i) \), and \( \sigma_i \) is the (true) standard deviation of the response of design \( i \). However, in reality the value of \( \sigma_i \) is unknown, so the sample standard deviations \( S_i(N_i) \) will be used. Moreover, the two ratios should stay the same when the objective is to minimize the computing budget with a pre-specified minimal CS probability \( P^* \).

2.4 An Enhanced Two-Stage Selection (ETSS) Procedure

Chen and Kelton (2000) proposed an ETSS procedure that takes into account not only the sample variances but also the difference between sample means across designs. They show that the ETSS procedure significantly reduces the number of total simulation replications or batches compared to Rinott’s procedure. However, ETSS is a heuristic procedure and does not guarantee \( P(CS) \geq P^* \).

Let \( \bar{X}_b(n_0) = \min_{i=1}^{k} \bar{X}_i(n_0) \) and

\[ d_i = \max(d^*, \bar{X}_i(n_0) - \bar{X}_b(n_0)). \] \hspace{1cm} (4)

ETSS computes the number of required simulation replications or batches for each design as

\[ N_i = \max(n_0, \lceil (h S_i(n_0)/d_i)^2 \rceil), \quad \text{for } i = 1, 2, \ldots, k. \] \hspace{1cm} (5)

The difference between equations (5) and (1) is that \( d_i \) instead of \( d^* \) is used. The differences of the sample means are embedded in \( d_i \), so this procedure will allocate fewer replications or batches to less promising designs \( i \) whose sample means \( \bar{X}_i(n_0) \) are far in excess of \( \bar{X}_b(n_0) \).

If \( N_i > n_0 \) and \( \bar{X}_i(n_0) - \bar{X}_i(n_0) \geq d^*, \) for \( l = 2, 3, \ldots, k \), the ratio

\[ \frac{N_i}{N_{i_2}} = \left( \frac{d_i}{d^*} \right)^2 \left( \frac{S_i(n_0)}{S_{i_2}(n_0)} \right)^2 \] \hspace{1cm} (6)

is the same as that in the OCBA (Chen et al. 2000). On the other hand, the ratio

\[ \frac{N_i}{N_{i_2}} = \left( \frac{d_i}{d^*} \right)^2 \left( \frac{S_i(n_0)}{S_{i_2}(n_0)} \right)^2 \] \hspace{1cm} (7)

is different from that of OCBA, because OCBA does not use the indifference parameter \( d^* \). If a design deviates less than \( d^* \) from the best design, OCBA will allocate extra samples to rank these two designs.
Generally speaking, we can improve the efficiency of R&S procedures with a pre-selection. The pre-selection approach is a screening device that attempts to select a (random-size) subset of the $k$ alternative designs that contains the best one. Inferior designs will be excluded from further consideration, reducing the overall simulation effort. In the ETSS procedure, design $i$, having the total required sample size $N_i = n_0$, can be viewed as being excluded from further consideration. Thus, the ETSS procedure has an intrinsic subset pre-selection built-in. Chen (2001) evaluates the performance of the built-in subset pre-selection in ETSS.

The ETSS procedure is derived based on $d_i = \max(d^*, \mu_i - \mu_{i_1})$; however, the true mean $\mu_i$ is unknown and $\bar{X}_i(n_0)$ from the first stage is used to estimate $d_i$, i.e., equation (4), which introduces noise into the procedure. Chen (2002) proposes adding a conservative adjustment to the ETSS procedures, which significantly improves $P(CS)$ by strategically allocating more replications or batches to more promising designs. Instead of using (4), the adjustment sets

$$d_i = \max(d^*, \bar{X}_i(n_0) - U(\bar{X}_b(n_0))),$$

where $U(\bar{X}_b(n_0))$ is the one-tailed upper $P^*$ confidence limit of $\mu_b$.

### 2.5 A fully Sequential Procedure for Indifference-Zone Selection

Kim and Nelson (2001) describe a fully sequential procedure, denoted as $KN$. The procedure is as follows.

1. **Setup**: Based on the input parameters: confidence level $P^*$, indifference-zone parameter $d^*$ and first-stage sample size $n_0$. calculate
   $$\eta = \frac{1}{2} \{[2(1 - (P^*)^{1/(k-1)})]^{-2/(n_0-1)} - 1\}.$$

2. **Initialization**: Let $I = \{1, 2, \ldots, k\}$ be the set of designs still in contention, and let $h^2 = 2\eta(n_0 - 1)$. Obtain $n_0$ observations $X_{ij}$, $j = 1, 2, \ldots, n_0$, from each design $i = 1, 2, \ldots, k$. For all $i \neq l$ compute
   $$S_{il}^2(n_0) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_{ij} - X_{lj} - [\bar{X}_i(n_0) - \bar{X}_l(n_0)])^2,$$
   the sample variance of the difference between design $i$ and $l$. Let
   $$N_{il} = \lceil (\frac{hS_{il}(n_0)}{d^*})^2 \rceil$$
where \( \lfloor z \rfloor \) is the greatest integer less than or equal to the real number \( z \), and let

\[
N_i = \max_{i \neq i} N_i.
\]

Here \( N_i + 1 \) is the maximum number of observations that can be taken from design \( i \). If \( n_0 \geq \max_i N_i + 1 \) then stop and select the design with the smallest \( \bar{X}_i(n_0) \) as the best. Otherwise set the observation counter \( t = n_0 \) and go to Screening.

3. **Screening**: Set \( I^{old} = I \). Let

\[
I = \{ i : i \in I^{old} \text{ and } \bar{X}_i(r) \leq \bar{X}_i(r) + W_{il}(r), \forall l \in I^{old}, l \neq i \}
\]

where

\[
W_{il}(r) = \max \left\{ 0, \frac{d^r}{2r} \left( \frac{h S_{il}(r)}{d^r} \right)^2 - r \right\}.
\] (8)

4. **Stopping Rule**: If \( |I| = 1 \), then stop and select the design whose index is in \( I \) as the best. Here \( |I| \) is the cardinality of the set \( I \). Otherwise, take one additional observational observation \( X_{i,r+1} \) from each design \( i \in I \) and set \( r = r + 1 \). If \( r = \max_i N_i + 1 \), then stop and select the design whose index is in \( I \) and has the smallest \( \bar{X}_i(r) \) as the best. Otherwise go to Screening.

### 2.6 Motivation

It is well known that R&S procedures developed based on the LFC are conservative, and two-stage procedures rely heavily on information obtained from just one stage. OCBA is a sequential procedure that takes into consideration the difference between mean estimators. However, OCBA is a fixed-sample-size procedure; it attempts to maximize \( P(\text{CS}) \) with a fixed total computing budget. Moreover, the achieved \( P(\text{CS}) \) cannot be estimated until the end of the procedure. If the achieved \( P(\text{CS}) \) is less than desired then the simulation needs to be restarted with a bigger simulation budget. The KN procedure described in Kim and Nelson (2001) requires all pairwise comparisons at each iteration, which is computationally intensive especially when the number of alternatives under consideration is large. If \( k' = |I| \) at the \( j^{th} \) iteration, the KN procedure requires \( k'(k' - 1)/2 \) pairwise comparisons at the \( j^{th} \) iteration. On the other hand, all other procedures described in this paper are basically based on multiple comparisons with a control, design \( i_1 \) being the control, and require only \( k' - 1 \) comparisons at the \( j^{th} \) iteration. The ETSS procedure, a straightforward enhancement of Rinott’s procedure, also suffers the drawback of two-stage procedures. To eliminate the drawback of relying too
heavily on the information from only one stage, we develop sequential R&S procedures from
the ETSS procedure.

3. METHODOLOGIES

In this section, we first present a simple sequential R&S procedure, denoted as the SRS proce-
dure, based on the ETSS procedure. We then present several variants of the SRS procedure.

As with most R&S procedures, we require the input data to be i.i.d. normal. Many per-
formance measures of interest are taken over some average of a sample path or a batch of
samples. Thus, the simulation output tends to be normally distributed in many applications.

If the nonnormality of the samples is a concern, users can use batch means (see Law and
Kelton 2000) to obtain sample means that are essentially i.i.d. normal.

3.1 Controlled Distance

To facilitate the derivation of the P(CS) of the ETSS procedure, we assume we know the true
means and variances. Without lost of generality, assume $\mu_{i_1} + d^* \leq \mu_{i_2} \leq \ldots \leq \mu_{i_k}$. If the
null hypothesis to be tested is $H_0 : \mu_{i_l} \leq \mu_{i_1}$ and the alternative hypothesis is $H_1 : \mu_i > \mu_{i_1}$,
then the test statistic that will be used to make a decision whether or not to reject the null
hypothesis is

$$t = \frac{\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1})}{S_{\bar{X}_{i_l} - \bar{X}_{i_1}}},$$

where $S_{\bar{X}_{i_l} - \bar{X}_{i_1}}^2$ is the sample variance of the values $\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1})$. The rejection region
for the alternative hypothesis at $\alpha$ level is $t > t_{1-\alpha,f}$ (see Rice 1995), where $t_{1-\alpha,f}$ denotes the
$1 - \alpha$ quantile of the $t$ distribution with $f$ df (degrees of freedom); see Rice (1995) for how $f$
can be estimated.

We reject the null hypothesis only if $t > t_{1-\alpha,f}$, or similarly

$$\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) > t_{1-\alpha,f}S_{\bar{X}_{i_l} - \bar{X}_{i_1}} = w_{i_l}.$$

By definition, a one-tailed $1 - \alpha$ CI half-width $w_{i_l}$ ensures $P[\mu_{i_l} - \mu_{i_1} \geq \bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) - w_{i_l}] \geq 1 - \alpha$. Moreover, for us to conclude with confidence $1 - \alpha$ that $\mu_i > \mu_{i_1}$ the lower
endpoint of the one-tailed $1 - \alpha$ CI must be positive, i.e., $\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) - w_{i_l} > 0$.

For details on the duality of confidence intervals and hypothesis tests see Rice (1995). By
symmetry of the normal distribution, $P[\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) + w_{i_l} \geq \mu_{i_l} - \mu_{i_1}] \geq 1 - \alpha$, i.e.,
$P[\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) \geq (\mu_{i_l} - \mu_{i_1}) - w_{i_l}] \geq 1 - \alpha$. The half-width $w_{i_l}$ depends on the
sample sizes and becomes smaller as the sample sizes become larger. If the sample sizes $N_i$ and $N_j$ are large enough so that the one-tailed $1 - \alpha$ CI half-width $\mu_i - \mu_j > w_{ij}$, then \( P[\bar{X}_{ij}(N_i) - \bar{X}_{ij}(N_j) > 0] \geq 1 - \alpha. \)

Let event $E_l$, for $l = 2, 3, \ldots, k$, denote $\bar{X}_{ij}(N_i) - \bar{X}_{ij}(N_j) > 0$. With independent sampling across designs, the $E_l$’s are positively correlated, and by Slepian’s inequality (Tong 1980)

\[
P(\text{CS}) = P[E_l \text{ for } l = 2, 3, \ldots, k] \geq \prod_{l=2}^{k} P[E_l].
\]

The equality holds for $k = 2$, and for $k > 2$ the equation holds with strict inequality. Let $P_1 = (P^*)^{\frac{k-1}{2}}$. Rinott’s constant $h$ is obtained by solving $E(\prod_{l=2}^{k} P[E_l]) = P^*$; see Appendix A. Thus, after the constant $h$ is assigned a numeric value, it ensures $P[E_l] \geq P_1$ for $l = 2, 3, \ldots, k$.

Let $w_{ij}$ denote the one-tailed $P_1$ CI half-width. If the total sample size $\sum_{i=1}^{k} N_i$ is large enough so that $\mu_{ij} - \mu_{ij} > w_{ij}$, for $l = 2, 3, \ldots, k$, then $P(\text{CS}) \geq P^*$. We can use the difference between sample means to derive $P(\text{CS})$ without the LFC assumption. If the half-width $w_{ij}$ of the one-tailed $P_1$ CI of $\mu_{ij} - \mu_{ij}$ is less than the difference between the true means, we will have the same confidence that $\bar{X}_{ij}(N_i) > \bar{X}_{ij}(N_j)$, i.e., $P[\bar{X}_{ij}(N_i) > \bar{X}_{ij}(N_j)] \geq P_1$. This makes sense when our objective is to achieve $P[\bar{X}_{ij}(N_i) < \bar{X}_{ij}(N_j)]$ for $l = 2, 3, \ldots, k \geq P^*$. If the objective of the simulation experiments is to estimate the differences of the expected responses, we can use different experimental designs and different procedures to obtain more precise estimates.

Indifference-zone selection procedures seek to avoid allocating large sample sizes to rank designs $i$ with $\mu_i - \mu_{ij} < d^*$. One way to look at the indifference amount is that we would like to rank the designs with a desired confidence when their performance measures deviate more than $d^*$. More specifically, when $\mu_i \leq \mu_{ij} + d^*$, indifference-zone procedures cannot guarantee the order of these designs with a desired confidence. On the other hand, if $\mu_{ij} > \mu_{ij} + d^*$, indifference-zone procedures will guarantee the order of these designs with a desired confidence.

Let the controlled distance $d_{ij} = \max(d^*, \mu_{ij} - \mu_{ij})$. If design $i_2$ is a good design with mean $\mu_{ij} - \mu_{ij} < d^*$, then $d_{ij} = d^*$. Thus, ETSS has the ability to pick a good design with probability $P^*$ without allocating extra samples to rank designs $i_1$ and $i_2$. Since $d_{ij} \geq d^*$, smaller sample sizes are adequate to achieve half-width $w_i < d_i$ rather than $w_i < d^*$.

For example, if the number of designs under consideration is $k = 2$ and $\mu_{ij} + d^* < \mu_{ij}$, then a sample size that achieves $d^* > w_{ij}$ will guarantee that $P[\bar{X}_{ij}(N_{ij}) > \bar{X}_{ij}(N_{ij})] \geq P^*$. We know that the sample sizes determined by Rinott’s procedure guarantee

\[
P(\text{CS}) = P[\bar{X}_{ij}(N_{ij}) - \bar{X}_{ij}(N_{ij}) > 0] \geq P^*,
\]
and under the LFC
\[ P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) \geq (\mu_{i2} - \mu_{i1}) - d^* = 0] \geq P^*, \]
and
\[ P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) \geq (\mu_{i2} - \mu_{i1}) - w_{i2} \geq 0] \geq P^*. \]

On the other hand, the ETSS procedure attempts to guarantee
\[ P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) \geq (\mu_{i2} - \mu_{i1}) - d_{i2}] \geq P^*, \]
where \( d_{i2} = \max(d^*, \mu_{i2} - \mu_{i1}) \). If \( \mu_{i2} - \mu_{i1} \geq d^* \), then \( x = \mu_{i2} - \mu_{i1} - d_{i2} = 0 \), and
\[ P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) \geq 0] \geq P^*. \]
If \( \mu_{i2} - \mu_{i1} < d^* \), then \( x = \mu_{i2} - \mu_{i1} - d_{i2} < 0 \), and
\[ P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) \geq x] \geq P^*. \]

Therefore, \( P[\bar{X}_{i2}(N_{i2}) - \bar{X}_{i1}(N_{i1}) > 0] \) may be less than \( P^* \). However, if we pick design \( i_2 \), it will be considered as a correct selection by definition.

We therefore have the following result:

**Theorem 1** For \( k \) competing designs whose observations \( X_{ij} \) are normally distributed with means \( \mu_1, \mu_2, \ldots, \mu_k \) and unknown variances that need to be estimated by sample variances \( S_1^2(r), S_2^2(r), \ldots, S_k^2(r) \), where \( r \) is the current sample size, \( P(\text{CS}) \) will be at least \( P^* \) when the sample size for design \( i \) is
\[ N_i = \max(n_0, \lceil (hS_i(r)/d_i)^2 \rceil), \quad \text{for } i = 1, 2, \ldots, k, \]
where the critical value \( h \) is the same as in Rinott’s procedure and \( d_i = \max(d^*, \mu_i - \mu_{i1}) \).

**Proof:** See Appendix A. Note that this result and its proof assume knowledge of the true means.

That is, if we allocated \( n_i \) to design \( i \) and \( n_i \geq N_i \), then \( P(\text{CS}) \geq P^* \). Moreover, if \( n_i \) is increased sequentially, then design \( i \) can be excluded from further simulation as soon as \( n_i \geq N_i \). Since we do not know the true means, however, the ETSS procedure and its variants use sample means to estimate \( d_i \) so do not guarantee \( P(\text{CS}) \geq P^* \); hence, we conduct empirical experiments in Section 4 to evaluate the ETSS procedure.
3.2 A Sequentialized Ranking and Selection Procedure

With Theorem 1, we now present a cost-effective sequential approach to select the best design from \( k \) alternatives.

**The Sequential R&S algorithm:**

1. Initialize the set \( I \) to include all \( k \) designs. Simulate \( n_0 \) replications or batches for each design \( i \in I \). Set the iteration number \( j = 0 \), and \( N_{1,j} = N_{2,j} = \ldots = N_{k,j} = n_0 \), where \( N_{i,j} \) is the sample size allocated for design \( i \) at the \( j^{th} \) iteration.

2. Set \( j = j + 1 \) and compute \( \delta_{ij} \), the needed incremental number of replications or batches for design \( i \) at the \( j^{th} \) iteration according to equation (9).

3. If \( \delta_{ij} = 0 \) and \( i \neq b \) (where \( \bar{X}_{b,j} = \min_{i \in I} \bar{X}_{i,j} \)), delete design \( i \) from the subset \( I \).

4. If there is only one element (or the pre-determined number of best designs) in the subset \( I \) or we have exhausted the computing budget, go to step 6.

5. Simulate \( \delta_{ij} \) additional replications or batches for each design \( i \in I \) at the \( j^{th} \) iteration. Go to step 2.

6. Return the values \( b \) and \( \bar{X}_b \), where \( \bar{X}_b = \min_{1 \leq i \leq k} \bar{X}_i \).

We can reduce the number of iterations with a larger incremental sample size \( \delta_{ij} \) for design \( i \) at the \( j^{th} \) iteration, but we run the risk of allocating more samples than necessary to non-promising designs. We propose to compute \( \delta_{ij} \) dynamically with all the information obtained up to the current iteration. Let \( N_{i,j} \) denote the sample size allocated for design \( i \) at the \( j^{th} \) iteration, \( N_{i,0} = n_0 \), \( \bar{X}_{i,j} \) denote the sample mean of design \( i \) at the \( j^{th} \) iteration, \( \bar{X}_{b,j} \) denote the smallest sample mean at the \( j^{th} \) iteration, i.e., \( \bar{X}_{b,j} = \min_{i \in I} \bar{X}_{i,j} \), and

\[
d_{i,j} = \max(d^*, \bar{X}_{i,j} - \bar{X}_{b,j}).
\]

Since sample means instead of true means are used to compute \( d_{i,j} \), the proposed procedure does not guarantee \( P(CS) \geq P^* \), however, conservative users can use the adjusted ETSS (Chen 2002) to compute \( d_{i,j} \), i.e., set

\[
d_{i,j} = \max(d^*, \bar{X}_{i,j} - U(\bar{X}_{b,j}))
\]
to increase $P(CS)$. Furthermore, as the sequentialized R&S procedure proceeds, the precision of $d_{i,j}$ becomes better as larger sample sizes are used to obtain sample means.

In the proposed sequential procedure, all the alternatives $1 \leq i \leq k$ are included initially in the set $I$ for R&S, and the additional sample size for alternative $i$ at iteration $j + 1$ is

$$\delta_{i,j+1} = \lceil ((hS_i(N_{i,j})/d_{i,j})^2 - N_{i,j})^+/2 \rceil,$$

where $w^+ = \max(w, 0)$. We use the equation $S_i^2(r) = (\sum_{j=1}^{r} X_{ij}^2/r - \bar{X}_i^2)r/(r-1)$ to compute the variance estimator, so we are only required to store the triple $(N_{i,j}, \sum_{j=1}^{N_{i,j}} X_{ij}, \sum_{j=1}^{N_{i,j}} X_{ij}^2)$ instead of the entire sequences $(X_{i1}, X_{i2}, \ldots, X_{iN_{i,j}})$. The total sample size for design $i$ at iteration $j + 1$ is $N_{i,j+1} = N_{i,j} + \delta_{i,j+1}$. Design $i \neq b$ having $N_{i,j+1} = N_{i,j}$, i.e., $\delta_{i,j+1} = 0$, will be excluded from further consideration, i.e., design $i$ will be deleted from subset $I$. Design $b$ will not be excluded from further consideration even when $\delta_{b,j} = 0$ because it is likely to have the smallest sample mean in later iterations and can provide a better value of $d_{i,j}$. Thus, the SRS procedure can eliminate inferior designs at any stage (iteration) and reduce the overall computational effort required to find the best.

The SRS procedure is able to estimate the required sample size for each design to obtain the specified minimal $P(CS)$ based on information obtained up to the current stage, so we are able to allocate incremental sample sizes intelligently. The procedure allocates a large incremental sample size at the first iteration and reduces the incremental sample sizes approximately by half at later iterations; see (9). This makes sense because there should be room for more aggressive budget allocation at early iterations, which reduces the number of iterations and the overhead in computing sample means and sample variances. Furthermore, we don’t run the risk of allocating more samples than necessary because we have estimated the required sample size for each design. As the procedure proceeds, the incremental sample size allocation become less aggressive to avoid allocating more samples than necessary to non-promising designs.

Furthermore, the factor of 1/2 in (9) insures that the allocated sample sizes satisfy approximately the ratios of (6) and (7). Assuming $(hS_i(n_0)/d_i)^2 - n_0 > 0$ at the first iteration, then the sample size for each alternative $i N_{i,1}$ is a little more than half of $N_i$ obtained from (5). Therefore, the ratios of (6) and (7) should be roughly the same when $N_i$ is replaced by $N_{i,1}$. On the other hand, if the procedure stops when the subset contains only the estimated best design, i.e., $|I| = 1$, then sample size $N_{i,j}$ should be close to $N_i$. Thus, the ratios should also be roughly the same. Consequently, the required ratios of sample size for each alternative to maximize $P(CS)$ is maintained approximately at all iterations.
The critical value \( h \) depends on \( P^*, k, \) and \( n_0 \). Even though the number of designs under consideration and the sample sizes for each design will change at each iteration, we can use the initial value of \( h \) through all iterations. This simplifies the programming effort and provides conservative estimates of the sample sizes. On the other hand, the critical value can be updated at each iteration with \( k' = |I| \) and \( N_{i,j} \). Since \( h \) is provided in the tables only for several pre-determined sample sizes, we can use the largest sample size \( n_j \) in the tables that satisfies \( n_j \leq \min_{1 \leq i \leq k'}(N_{i,j}) \) at the \( j^{th} \) iteration to obtain \( h \). Alternatively, \( h \) can also be dynamically computed with \( k' \) and \( n_j \) at each iteration, where \( n_j = \min_{1 \leq i \leq k'}(N_{i,j}) \).

When we use \( k' \) instead of \( k \) to compute the critical value, it implies that design \( i \) that was eliminated from further simulation has \( P[\bar{X}_i > \bar{X}_b] = 1 \), which of course is not true. Therefore, \( P(\text{CS}) \) may be less than the specified \( P^* \). However, this method can be used when the number of alternatives \( k \) under consideration is large. Furthermore, we can set a minimal number \( k'' \) for computing the critical value \( h \). We will use \( k' \) to compute the critical value \( h \) when \( k' > k'' \) and use \( k'' \) when \( k' \leq k'' \). In this way, we can eliminate inferior designs at early iterations to reduce overall computational effort and ensure high \( P(\text{CS}) \) among the estimated best \( k'' \) designs.

The procedure will iterate repeatedly until we have exhausted the pre-determined computing budget or subset \( I \) contains only the pre-determined best \( b' \) alternatives, i.e., \( b' = |I| \), and \( 1 \leq b' \leq k \). The total incremental sample size at the \( j^{th} \) iteration is

\[
\Delta_j = \sum_{1 \leq i \leq k'} \delta_{i,j}.
\]

### 3.3 Discussion of the Procedure

The sample size allocation strategy of the SRS procedure is to maximize \( P(\text{CS}) \) with a given budget, i.e., the ratios of sample sizes satisfy (6) and (7). Since the allocated sample size for each design of the SRS procedure is approximately consistent with OCBA, the allocated sample sizes should be near optimal. However, performance of the SRS procedure can be further improved by all pairwise comparisons or multiple comparisons with the best \( g \) (\( 1 \leq g \leq k' \)) designs at each iteration. The number of iterations of the SRS procedure is much less than the KN procedure, so it will not be as computationally intensive. For example, if design \( b \), where \( \bar{X}_b = \min_{1 \leq i \leq k} \bar{X}_i \), has large variance and the second best design \( s \) has small variance, then the SRS procedure will not be able to eliminate certain inferior designs at early iterations. This is because the SRS procedure is based on multiple comparisons with the estimated best,
i.e., design $b$, which happens to have large variance. Thus, we cannot conclude with a desired confidence that $\bar{X}_i > \bar{X}_b$ for some design $i$. However, if we use design $s$ as a control to perform multiple comparisons, then we may be able to eliminate certain inferior designs because the variance of design $s$ is small and we will be able to conclude with a desired confidence that $\bar{X}_i > \bar{X}_s$ for some design $i$, i.e., we have concluded that $P[\bar{X}_i - \bar{X}_s] \geq P = 1 - \frac{1-P^*}{k-1}$ (see Proposition 1 below). Consequently, we may be able to reduce the overall computational effort to find the best.

To incorporate all pairwise comparisons and to improve the overall computing efficiency, we can use different incremental sample size allocation strategies depending on the required computing effort to obtain samples. A less aggressive sample size allocation strategy is as follows. Let

$$
\delta_{i,j+1} = \lceil (\frac{(hS_i(N_{i,j})}{d_{i,j}})^2 - N_{i,j}) + \rceil.
$$

If $i \neq b$ and $\delta_{i,j+1} = 0$, then remove design $i$ from subset $I$. We then use the same incremental sample size $\delta_j$ at the $j$th iteration for every design $i \in I$, where

$$
\delta_j = \min_{i \neq b, i \in I} \delta_{i,j}.
$$

Using this sample size increment rule, $N_{i,j}$’s are the same for all $i \in I$ at the $j$th iteration. We can then apply the screening step of the KN procedure (see Section 2.5) in the SRS procedure; we denote this modified procedure as SMC1 in the remainder of this paper. To avoid storing the entire sequences of observations, $S^2_{il}(r)$ is approximated by $S^2_i(r) + S^2_l(r)$. Another variation of this procedure is to use a modified two-sample-t confidence interval; see Law and Kelton (2000). Let $P = 1 - \frac{1-P^*}{k-1}$. We will eliminate design $l$ if

$$
\bar{X}_l(r) > \bar{X}_i(r) + t_{P,r-1} \sqrt{\frac{S_l(r)^2}{r} + \frac{S_i(r)^2}{r}}, \quad i, l \in I,
$$

i.e., the procedure has concluded that $P[\bar{X}_l(r) > \bar{X}_i(r)] \geq P$. This variation of the modified procedure is denoted as SMC2 and is summarized in the following proposition:

**Proposition 1.** Let the set $I$ contain $k$ competing designs whose $X_{ij}$ are normally distributed with unknown means and unknown variances that need to be estimated by sample means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ and sample variances $S^2_1(r), S^2_2(r), \ldots, S^2_k(r)$. If $k - 1$ designs are removed (eliminated) sequentially from $I$ such that each eliminated design $j$ satisfies the $\bar{X}_j > \bar{X}_i + w_{ij}, \quad i, j \in I$ (where $w_{ij} = t_{P,r-1} \sqrt{\frac{S_i(r)^2}{r} + \frac{S_j(r)^2}{r}}$ and $P = 1 - (1 - P^*/(k - 1))$, then $P[i_1 \in I] \geq P^*$. Note that the sample size $r$ changes from iteration to iteration, but $r$ is regarded as a constant at any particular iteration.
Proof: Note that $w_{ij}$ as defined in the proposition is the one-tailed level $1 - \alpha$ CI half-width for $\mu_i - \mu_j$. Appendix B shows that $P[\bar{X}_i > \bar{X}_j + w_{ij}] \leq \alpha$. Let $\alpha = (1 - P^*)/(k - 1)$ and $P(ICS_i)$ denote the probability of incorrect selection in each hypothesis test; then $P(ICS_i) = (1 - P^*)/(k - 1)$.

When the modified two-sample-t test is applied to all $k$ designs, we follow Corollary 1 of Kim and Nelson (2001) to show that

$$P[i_1 \in I] = P[\text{designs } i_2, i_3, \ldots, i_k \text{ are eliminated}]$$
$$\geq P[\text{design } i_1 \text{ eliminates } i_2, i_3, \ldots, i_k]$$
$$= 1 - P[ICS; \text{ one of } i_2, i_3, \ldots, k \text{ eliminates } i_1]$$
$$\geq 1 - \sum_{i=1}^{k-1} P[ICS_i]$$
$$= P^*$$

If the objective is to maximize $P(\text{CS})$ with a given computing budget and the user also specifies the required minimal $P(\text{CS})$, then we can use the sample size allocation strategy of (9) and iterate until the pre-determined computing budget $T$ is exhausted. However, if the user does not specify the required minimal $P(\text{CS})$ or the value of $h$ is not available, we can use the following rule:

$$\Delta_l = \min(T, \max(k, \lceil T/2 \rceil))$$

Then $\Delta_l$ is distributed to each design so that the $N_{i,l}$’s satisfy the ratio of (6) and (7) approximately; see Chen et al. (2004), which compares the performance of OCBA and a procedure that has allocated sample sizes satisfying the ratios derived from the ETSS procedure. They conclude that, with a fixed computing budget, the difference of the observed $P(\text{CS})$’s from these two procedures is minor.

One drawback is that without the required minimal $P(\text{CS})$, we will not be able to exclude inferior designs from consideration as the simulation proceeds. In both rules, the sequential procedure allocates incremental sample sizes aggressively at earlier iterations and becomes less aggressive as the procedure proceeds and the allocated sample sizes $N_{i,j}$ are close to the required sample size $N_i$. This way we will be able to reduce the number of iterations without the risk of putting too much effort into simulating non-promising designs.

The ETSS procedure uses the difference between sample means and sample variances to compute the required sample size for each design. We reduce the variance of the difference be-
tween sample means when we use common random numbers (CRN) with the ETSS procedure, so we improve the P(CS) of the procedure. Furthermore, if we use
\[ P = (1 - \frac{1 - P^*}{k-1})^{k-1} \]
instead of \( P^* \) to compute the constant \( h \) so that
\[ P[E_l] \geq 1 - \frac{1 - P^*}{k-1}, \text{ for } l = 2, 3, \ldots, k, \]
(see Section 3.1) then by the Bonferroni inequality (Law and Kelton 2000),
\[ P(CS) = P[E_l, \text{ for } l = 2, 3, \ldots, k] \geq 1 - \sum_{l=2}^{k}(1 - P[E_l]) \geq P^*. \]
Hence, we can use CRN to increase the P(CS) of R&S procedures without any further assumption. We recommend using CRN with the sequential R&S procedures as well. Because CRN can reduce the variance of the difference between sample means, it would improve the precision of pairwise comparisons and allow inferior designs to be eliminated earlier in the iterations.

The KN procedure is very effective at finding the best design but it is computationally intensive not only because it requires all pairwise comparisons, but also because it increases only one additional replication for each design in contention at each iteration. Thus, it generally requires many more iterations than do our procedures. The KN procedure also estimates the required number of total simulation replications \( N_i \) for each design \( i \in I \) at the beginning of the procedure. However, the ratio \( N_i/N_j \) determined by the KN procedure does not maximize P(CS), so the KN procedure cannot use the same scheme we proposed for our procedures to allocate additional replications.

4. **EMPIRICAL EXPERIMENTS**

In this section we present some empirical simulation results of the Rinott, ETSS, KN, SRS, SMC1, and SMC2 procedures. We use the critical value \( h \) obtained from tables with the required minimal probability of correct selection \( P^* \), the number of designs under consideration \( k \), and the sample size at first stage \( n_0 \) for all iterations. Because the critical value \( h \) is not updated at later iterations with larger sample sizes, the SRS, SMC1, and SMC2 procedures should obtain higher P(CS) with larger sample sizes than necessary. Instead of using stochastic systems simulation examples, which offer less control over the factors that affect the performance of a procedure, we use various normally distributed random variables to represent the systems.
Table 1: $\hat{P}(CS)$ and Sample Sizes for Experiment 1

<table>
<thead>
<tr>
<th>Procedure</th>
<th>$P^* = 0.90$</th>
<th>$P^* = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{P}(CS)$</td>
<td>$T$</td>
</tr>
<tr>
<td>Rinott(20)</td>
<td>0.9934</td>
<td>5256</td>
</tr>
<tr>
<td>ETSS(20)</td>
<td>0.9508</td>
<td>1197</td>
</tr>
<tr>
<td>KN(20)</td>
<td>0.9981</td>
<td>1071</td>
</tr>
<tr>
<td>SRS(20)</td>
<td>0.9925</td>
<td>1243</td>
</tr>
<tr>
<td>SMC1(20)</td>
<td>0.9876</td>
<td>834</td>
</tr>
<tr>
<td>SMC2(20)</td>
<td>0.9857</td>
<td>886</td>
</tr>
<tr>
<td>Rinott(30)</td>
<td>0.9942</td>
<td>5001</td>
</tr>
<tr>
<td>ETSS(30)</td>
<td>0.9689</td>
<td>1234</td>
</tr>
<tr>
<td>KN(30)</td>
<td>0.9954</td>
<td>998</td>
</tr>
<tr>
<td>SRS(30)</td>
<td>0.9946</td>
<td>1221</td>
</tr>
<tr>
<td>SMC1(30)</td>
<td>0.9905</td>
<td>896</td>
</tr>
<tr>
<td>SMC2(30)</td>
<td>0.9901</td>
<td>980</td>
</tr>
</tbody>
</table>

4.1 Experiment 1: Equal Variances

There are ten alternative designs in the selection subset. Suppose $X_{ij} \sim \mathcal{N}(i, \sigma^2)$, $i = 1, 2, \ldots, 10$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. We want to select a design with the minimum mean: design 1. The indifference amount $d^*$ is set to 0.90 for all cases. We compare the actual $P(CS)$ of the Rinott, ETSS, KN, SRS, SMC1, and SMC2 procedures. We use two different initial number of replications, $n_0 = 20$ and 30. Furthermore, 10,000 independent experiments are performed to estimate the actual $P(CS)$ by $\hat{P}(CS)$: the proportion of the 10,000 experiments in which we obtained the correct selection.

The results of our experiment 1 are summarized in Table 1. The $\hat{P}(CS)$ column lists the proportion of correct selection. The $T$ column lists the average of the total number of simulation replications ($T = \sum_{R=1}^{10000} \sum_{i=1}^{10} T_{R,i}/10000$, and $T_{R,i}$ is the number of total replications or batches for design $i$ at the $R$th simulation run) used in each procedure. The std($T$) column lists the standard deviation of the number of total sample sizes $T$ at each independent simulation run. The Rinott(20), ETSS(20), KN(20), SRS(20), SMC1(20), and SMC2(20) rows list the results of the respective procedures with initial replications $n_0 = 20$ (and similarly for $n_0 = 30$).

The $\hat{P}(CS)$’s are all larger than the specified $P^*$. SRS has better $\hat{P}(CS)$ than the ETSS procedure with no significant increase in the total number of replications. Because the variance of the sample is larger with a smaller initial sample size $n_0$, the SRS procedure allocates more
Table 2: Detailed Sample Sizes for $P^* = 0.90$ and $n_0 = 20$ of Experiment 1

<table>
<thead>
<tr>
<th>Design</th>
<th>Rinott</th>
<th>ETSS</th>
<th>KN</th>
<th>SRS</th>
<th>SMC1</th>
<th>SMC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>526</td>
<td>458</td>
<td>258</td>
<td>495</td>
<td>303</td>
<td>325</td>
</tr>
<tr>
<td>2</td>
<td>527</td>
<td>304</td>
<td>255</td>
<td>366</td>
<td>242</td>
<td>281</td>
</tr>
<tr>
<td>3</td>
<td>526</td>
<td>175</td>
<td>148</td>
<td>156</td>
<td>102</td>
<td>99</td>
</tr>
<tr>
<td>4</td>
<td>526</td>
<td>96</td>
<td>75</td>
<td>49</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>522</td>
<td>48</td>
<td>78</td>
<td>40</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>524</td>
<td>30</td>
<td>62</td>
<td>26</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>522</td>
<td>22</td>
<td>51</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>526</td>
<td>20</td>
<td>43</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>526</td>
<td>20</td>
<td>37</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>527</td>
<td>20</td>
<td>32</td>
<td>20</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

samples with smaller $n_0$. The SRS achieves more improvement in $P(\text{CS})$ when $n_0$ is small. The observed $P(\text{CS})$’s of both SMC1 and SMC2 are slightly less than SRS, but the sample sizes are significantly smaller. Moreover, SMC1 has better performance than SMC2, achieving better $P(\text{CS})$ with less effort.

Table 2 lists the detailed simulation replications used for each design under different selection procedures with $n_0 = 20$ and $P^* = 0.90$. We do not list the results for $n_0 = 20$ and $P^* = 0.95$, and for $n_0 = 30$, because they are similar to Table 2. The Rinott, ETSS, KN, SRS, SMC1, and SMC2 columns list the average number of simulation replications for each design and each procedure. Rinott’s procedure will be the same as equal allocation for additional simulation replications in this setting, i.e., the variances are equal for all designs. Our experimental results confirm that. On the other hand, in the ETSS-based procedures the number of additional simulation replications decreases as the differences $d_{i,b} = \bar{X}_i - \bar{X}_b (> 0)$ increase. This makes sense because as $\mu_i - \mu_b$ increases, it is more likely that $\bar{X}_i > \bar{X}_b$. In other words, as the observed difference between sample means across alternatives $d_{i,b}$ increases, it is less likely we would conclude that $\mu_i < \mu_b$.

The ratio of the average number of simulation replications allocated for design 10 and design 1 of Rinott’s procedure is $527/526 = 1.0019$ when $P^* = 0.90$ and $n_0 = 20$, which is close to the theoretical value $1 ((S_{10}(n_0)/S_1(n_0))^2 = (6/6)^2)$. On the other hand, this ratio is only $20/458 = 0.0437$ under the ETSS procedure. This is where ETSS-based procedures can significantly improve the efficiency of the Rinott procedure. Note that 0.0437 is much larger than the theoretical value $0.01 ((d^*/d_{i10})^2(S_{10}(n_0)/S_1(n_0))^2 = (0.9/9)^2)$ because $n_0 > N_i$ for some design $i$. For example, if we use $n_0 = 5 (\geq 0.01 \times 458)$, the ETSS procedure would have
eliminated design 10 at the initial iteration.

In this setting, all KN experiments terminated by the stopping rule that there is only one design in the subset $I$. This implies that the bound on the sample size determined by the procedure is conservative. However, by using all pairwise comparisons at each iteration, the KN procedure is able to achieve high precision with relatively little data. Unfortunately, it is computationally intensive to perform all pairwise comparisons at each iteration, especially when the inferior designs have not been eliminated at early iterations. For example, the average number of replications for design 10 is 32 when $P^* = 0.90$ and $n_0 = 20$ (see table 2); this means that on average, design 10 was not eliminated until the 13th iteration. On the other hand, the SRS family of R&S procedures is able to eliminate design 10 at the first stage. However, as the sample size $r$ becomes larger, the KN procedure becomes more aggressive in eliminating inferior designs because the larger the sample size $r$, the smaller the half width $W_{il}(r)$; see (8).

As indicated in Section 2.4, the efficiency of R&S procedures can be improved with subset pre-selection. The performance improvement of combining subset pre-selection and Rinott’s procedure is well known and often denoted as Rinott+Subset or Rinott+Screening. We did not include this procedure in our comparison because its results can be seen indirectly. Based on the experimental results of Chen (2001), we know which designs would likely be included in the subset from the results of the detailed sample size at each design. For example, designs 1 through 7, having $N_i > n_0$ in ETSS, are likely be included in the subset (see Table 2). Therefore, the average sample sizes for Rinott+Subset with $P^* = 0.90$ and $n_0 = 20$ will approximately be 3737 ($526 + 527 + \ldots + 522 + 20 + 20 + 20$).

4.2 Experiment 2: Increasing Variances

This is a variation of experiment 1. All settings are preserved except that the variance of each design increases as the mean increases. Namely, $X_{ij} \sim N(i, (6 + (i - 1)/2)^2)$, $i = 1, 2, \ldots, 10$.

The results are in Tables 3 and 4. Because most designs have larger variances than those in experiment 1, the total simulation effort is more than in experiment 1. We are less confident of the best selection with this setting. Therefore, more simulation replications are needed to obtain the desired confidence. In general, the implication of these results are similar to experiment 1. The observed $\hat{P}(CS)$’s are all larger than the specified $P^* = 0.90$ and $P^* = 0.95$. SRS does better than the ETSS procedure with no significant increase in the total number of replications. The observed $\hat{P}(CS)$’s of both SMC1 and SMC2 are slightly less than SRS.
Table 3: \( \hat{P}(CS) \) and Sample Sizes for Experiment 2

<table>
<thead>
<tr>
<th>Procedure</th>
<th>( \hat{P}(CS) )</th>
<th>( T )</th>
<th>std(( T ))</th>
<th>( \hat{P}(CS) )</th>
<th>( T )</th>
<th>std(( T ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rinott(20)</td>
<td>0.9930</td>
<td>10224</td>
<td>1101</td>
<td>0.9978</td>
<td>13020</td>
<td>1392</td>
</tr>
<tr>
<td>ETSS(20)</td>
<td>0.9426</td>
<td>1532</td>
<td>689</td>
<td>0.9534</td>
<td>1940</td>
<td>887</td>
</tr>
<tr>
<td>KN(20)</td>
<td>0.9981</td>
<td>1403</td>
<td>269</td>
<td>0.9996</td>
<td>1751</td>
<td>300</td>
</tr>
<tr>
<td>SRS(20)</td>
<td>0.9903</td>
<td>1539</td>
<td>379</td>
<td>0.9972</td>
<td>1969</td>
<td>460</td>
</tr>
<tr>
<td>SMC1(20)</td>
<td>0.9818</td>
<td>944</td>
<td>323</td>
<td>0.9911</td>
<td>1158</td>
<td>385</td>
</tr>
<tr>
<td>SMC2(20)</td>
<td>0.9773</td>
<td>886</td>
<td>412</td>
<td>0.9898</td>
<td>1114</td>
<td>505</td>
</tr>
<tr>
<td>Rinott(30)</td>
<td>0.9947</td>
<td>9757</td>
<td>864</td>
<td>0.9974</td>
<td>12280</td>
<td>1071</td>
</tr>
<tr>
<td>ETSS(30)</td>
<td>0.9636</td>
<td>1493</td>
<td>589</td>
<td>0.9709</td>
<td>1831</td>
<td>732</td>
</tr>
<tr>
<td>KN(30)</td>
<td>0.9969</td>
<td>1306</td>
<td>256</td>
<td>0.9992</td>
<td>1600</td>
<td>287</td>
</tr>
<tr>
<td>SRS(30)</td>
<td>0.9918</td>
<td>1453</td>
<td>335</td>
<td>0.9967</td>
<td>1818</td>
<td>390</td>
</tr>
<tr>
<td>SMC1(30)</td>
<td>0.9871</td>
<td>975</td>
<td>301</td>
<td>0.9936</td>
<td>1168</td>
<td>354</td>
</tr>
<tr>
<td>SMC2(30)</td>
<td>0.9868</td>
<td>988</td>
<td>386</td>
<td>0.9938</td>
<td>1176</td>
<td>473</td>
</tr>
</tbody>
</table>

Table 4: Detailed Sample Sizes for \( P^* = 0.90 \) and \( n_0 = 20 \) of Experiment 2

<table>
<thead>
<tr>
<th>Design</th>
<th>Rinott</th>
<th>ETSS</th>
<th>KN</th>
<th>SRS</th>
<th>SMC1</th>
<th>SMC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>526</td>
<td>447</td>
<td>281</td>
<td>504</td>
<td>304</td>
<td>292</td>
</tr>
<tr>
<td>2</td>
<td>619</td>
<td>348</td>
<td>275</td>
<td>430</td>
<td>259</td>
<td>276</td>
</tr>
<tr>
<td>3</td>
<td>717</td>
<td>247</td>
<td>175</td>
<td>217</td>
<td>127</td>
<td>104</td>
</tr>
<tr>
<td>4</td>
<td>822</td>
<td>164</td>
<td>134</td>
<td>126</td>
<td>69</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>932</td>
<td>107</td>
<td>112</td>
<td>80</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>1053</td>
<td>71</td>
<td>98</td>
<td>54</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>1178</td>
<td>49</td>
<td>89</td>
<td>40</td>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>1315</td>
<td>38</td>
<td>83</td>
<td>31</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td>9</td>
<td>1457</td>
<td>30</td>
<td>78</td>
<td>27</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>1601</td>
<td>27</td>
<td>75</td>
<td>24</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>
Table 5: $\hat{P}(\text{CS})$ and Sample Sizes for Experiment 3

<table>
<thead>
<tr>
<th>Procedure</th>
<th>$P^* = 0.90$</th>
<th>$P^* = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rinott(20)</td>
<td>0.9948 2361 293</td>
<td>0.9966 2989 367</td>
</tr>
<tr>
<td>ETSS(20)</td>
<td>0.9578 1038 339</td>
<td>0.9669 1279 429</td>
</tr>
<tr>
<td>KN(20)</td>
<td>0.9978 842 206</td>
<td>0.9992 1040 228</td>
</tr>
<tr>
<td>SRS(20)</td>
<td>0.9919 1063 214</td>
<td>0.9970 1338 245</td>
</tr>
<tr>
<td>SMC1(20)</td>
<td>0.9898 765 240</td>
<td>0.9953 914 291</td>
</tr>
<tr>
<td>SMC2(20)</td>
<td>0.9906 834 305</td>
<td>0.9934 1021 383</td>
</tr>
<tr>
<td>Rinott(30)</td>
<td>0.9933 2247 228</td>
<td>0.9966 2830 280</td>
</tr>
<tr>
<td>ETSS(30)</td>
<td>0.9749 1078 293</td>
<td>0.9792 1302 372</td>
</tr>
<tr>
<td>KN(30)</td>
<td>0.9964 825 191</td>
<td>0.9986 989 217</td>
</tr>
<tr>
<td>SRS(30)</td>
<td>0.9943 1069 209</td>
<td>0.9978 1316 238</td>
</tr>
<tr>
<td>SMC1(30)</td>
<td>0.9917 824 213</td>
<td>0.9973 967 259</td>
</tr>
<tr>
<td>SMC2(30)</td>
<td>0.9903 912 273</td>
<td>0.9959 1083 342</td>
</tr>
</tbody>
</table>

However, the sample sizes are significantly smaller. SMC1 has better performance than SMC2 in terms of $\hat{P}(\text{CS})$, sometimes even with less effort. All procedures allocate more additional simulation replications for designs with larger variances, as expected. For Rinott’s procedure the sampling allocation is based entirely on the variances, so $N_i > N_j$ when $S_i(n_0) > S_j(n_0)$. The ETSS-based procedures take into consideration the difference of the sample means, so $N_i < N_j$ even though $S_i(n_0) > S_j(n_0)$. The ETSS procedure has the most significant reduction in the number of replications or batches relative to Rinott’s procedure in this setting, i.e., the inferior alternatives have the largest variances.

Again, all KN experiments terminated with only one design in the subset $I$. Furthermore, the ability of the KN procedure to eliminate the inferior designs early worsens in this setting. For example, when $P^* = 0.90$ and $n_0 = 20$, on average the SRS family of procedures are able to eliminate design 10 within the first four iterations, but KN did not eliminate design 10 until the 56th iteration (see table 4).

### 4.3 Experiment 3: Decreasing Variances

This is another variation of experiment 1. All settings are preserved except that the variance of each design decreases as the mean increases. Namely, $X_{ij} \sim \mathcal{N}(i, (6 - (i - 1)/2)^2), i = 1, 2, \ldots, 10.$

The results are in Tables 5 and 6. Because most designs have smaller variances than those in experiment 1, the total simulation effort is less than in experiment 1. We are more confident
of the best selection with this setting. Therefore, fewer simulation replications are needed to obtain the desired confidence. All procedures allocate fewer additional simulation replications to inferior designs in this setting, i.e., as the sample means increase the variances decrease. Once again, SRS does better than ETSS with little additional data. The observed $\hat{P}(CS)$’s of both SMC1 and SMC2 are slightly lower than SRS, but the sample sizes are significantly smaller. SMC1 has better performance than SMC2, achieving better $\hat{P}(CS)$ with less effort. A handful (6 out of 40000) of these KN experiments terminated with $r = \max N_i + 1$ in this setting, but the rest terminated with $|I| = 1$. The KN procedure does better in eliminating inferior designs early in this setting.

### 4.4 Experiment: 4 LFC

In this experiment, we consider the least favorable configuration (LFC), $\mu_1 + d^* = \mu_2 = \ldots = \mu_{10}$. The minimum $P(CS)$ should occur at this configuration, where $\mu_1 = 0$ and $d^* = 1$. The variances are $\sigma_i^2 = 6^2$ for $i = 1, 2, \ldots, k$. We compare the $\hat{P}(CS)$’s of those procedures using two different initial replications $n_0 = 20$ and 30.

The results are in Table 7. Rinott’s procedure is generally conservative, but when we encounter the LFC, it is very effective. In this setting, i.e., with large variances and the LFC, the observed $\hat{P}(CS)$’s of the ETSS are less than the specified $P^*$. We don’t think this is a major drawback for the ETSS procedure because we rarely encounter LFC and it may not be too costly to select a design with the performance measure equals to or just greater than $\mu_{i_1} + d^*$. All the sequentialized R&S procedures have $\hat{P}(CS)$’s higher than the specified $P^*$. In this LFC setting, the KN procedure on average requires more than 201 iterations

---

Table 6: Detailed Sample Sizes for $P^* = .90$ and $n_0 = 20$ of Experiment 3

<table>
<thead>
<tr>
<th>Design</th>
<th>Rinott</th>
<th>ETSS</th>
<th>KN</th>
<th>SRS</th>
<th>SMC1</th>
<th>SMC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>525</td>
<td>470</td>
<td>236</td>
<td>486</td>
<td>303</td>
<td>337</td>
</tr>
<tr>
<td>2</td>
<td>443</td>
<td>267</td>
<td>234</td>
<td>308</td>
<td>224</td>
<td>258</td>
</tr>
<tr>
<td>3</td>
<td>367</td>
<td>125</td>
<td>122</td>
<td>104</td>
<td>80</td>
<td>82</td>
</tr>
<tr>
<td>4</td>
<td>297</td>
<td>49</td>
<td>76</td>
<td>40</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>233</td>
<td>24</td>
<td>50</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>179</td>
<td>20</td>
<td>35</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>131</td>
<td>20</td>
<td>26</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>91</td>
<td>20</td>
<td>21</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>59</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
Table 7: $\hat{P}(CS)$ and Sample Sizes for Experiment 4

<table>
<thead>
<tr>
<th>Procedure</th>
<th>$\hat{P}(CS)$</th>
<th>$T$</th>
<th>std($T$)</th>
<th>$\hat{P}(CS)$</th>
<th>$T$</th>
<th>std($T$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rinott(20)</td>
<td>0.9312</td>
<td>5268</td>
<td>544</td>
<td>0.9658</td>
<td>6685</td>
<td>680</td>
</tr>
<tr>
<td>ETSS(20)</td>
<td>0.6305</td>
<td>2070</td>
<td>845</td>
<td>0.6671</td>
<td>2641</td>
<td>1067</td>
</tr>
<tr>
<td>KN(20)</td>
<td>0.9603</td>
<td>2211</td>
<td>512</td>
<td>0.9866</td>
<td>2787</td>
<td>621</td>
</tr>
<tr>
<td>SRS(20)</td>
<td>0.9261</td>
<td>2701</td>
<td>879</td>
<td>0.9634</td>
<td>3743</td>
<td>1107</td>
</tr>
<tr>
<td>SMC1(20)</td>
<td>0.9162</td>
<td>2123</td>
<td>720</td>
<td>0.9625</td>
<td>2861</td>
<td>929</td>
</tr>
<tr>
<td>SMC2(20)</td>
<td>0.9201</td>
<td>2528</td>
<td>889</td>
<td>0.9607</td>
<td>3512</td>
<td>1143</td>
</tr>
<tr>
<td>Rinott(30)</td>
<td>0.9280</td>
<td>5004</td>
<td>414</td>
<td>0.9618</td>
<td>6305</td>
<td>523</td>
</tr>
<tr>
<td>ETSS(30)</td>
<td>0.7179</td>
<td>2294</td>
<td>832</td>
<td>0.7567</td>
<td>2885</td>
<td>1052</td>
</tr>
<tr>
<td>KN(20)</td>
<td>0.9511</td>
<td>2046</td>
<td>477</td>
<td>0.9791</td>
<td>2542</td>
<td>581</td>
</tr>
<tr>
<td>SRS(20)</td>
<td>0.9297</td>
<td>2635</td>
<td>826</td>
<td>0.9591</td>
<td>3579</td>
<td>1038</td>
</tr>
<tr>
<td>SMC1(20)</td>
<td>0.9221</td>
<td>2073</td>
<td>670</td>
<td>0.9651</td>
<td>2707</td>
<td>863</td>
</tr>
<tr>
<td>SMC2(20)</td>
<td>0.9194</td>
<td>2543</td>
<td>841</td>
<td>0.9649</td>
<td>3459</td>
<td>1075</td>
</tr>
</tbody>
</table>

and it took more than 2.5 hours to perform these 10000 simulation experiments on a Dell Latitude Laptop with 524K RAM. Since the incremental sample size is one, the number of iterations is approximately $T/k - n_0$. For example, the number of iterations is approximately $2211/10 - 20 \approx 201$ when $P^* = 0.90$ and $n_0 = 20$. On the other hand, SRS, SMC1, and SMC2 procedures on average require just several iterations and it took no more then 3 minutes to perform these same experiments. This is because under the LFC, the required sample sizes $N_i$ for designs $i = 2, 3, \ldots 10$ will be the same and the first incremental sample size $\delta_1$ will be close to $N_i - n_0$.

4.5 Experiment 5: SMC1 and SMC2 with Different Incremental Sample Sizes

In order to compare the SMC1, SMC2, and KN procedures in a similar setting, the incremental sample size $\delta_{i,j}$ for a SMC1 and SMC2 is set to 1 at each iteration in this experiment. We also tested another variation of the procedure, SMC3, which is in the same spirit as KN. We compute an upper bound on the sample size and use all pairwise comparisons to eliminate inferior designs as the procedure proceeds. Let

$$N = \max_{1 \leq i \leq k} N_i,$$

where $N_i$ is computed from (5) from the initial iteration. We will eliminate inferior designs from subset $I$ by (11). The SMC3 procedure will stop when $|I| = 1$ or $r > N$. That is,
Table 8: $\hat{P}(CS)$ and Sample Sizes for Experiment 5

<table>
<thead>
<tr>
<th>Set</th>
<th>Procedure</th>
<th>$P^* = 0.90$</th>
<th>$P^* = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{P}(CS)$</td>
<td>$T$</td>
<td>std($T$)</td>
</tr>
<tr>
<td>1</td>
<td>SMC1(20)</td>
<td>0.9865</td>
<td>685</td>
</tr>
<tr>
<td></td>
<td>SMC2(20)</td>
<td>0.9858</td>
<td>752</td>
</tr>
<tr>
<td></td>
<td>SMC3(20)</td>
<td>0.9887</td>
<td>786</td>
</tr>
<tr>
<td></td>
<td>SMC1(30)</td>
<td>0.9874</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>SMC2(30)</td>
<td>0.9892</td>
<td>818</td>
</tr>
<tr>
<td></td>
<td>SMC3(30)</td>
<td>0.9898</td>
<td>857</td>
</tr>
<tr>
<td>2</td>
<td>SMC1(20)</td>
<td>0.9783</td>
<td>833</td>
</tr>
<tr>
<td></td>
<td>SMC2(20)</td>
<td>0.9774</td>
<td>840</td>
</tr>
<tr>
<td></td>
<td>SMC3(20)</td>
<td>0.9868</td>
<td>886</td>
</tr>
<tr>
<td></td>
<td>SMC1(30)</td>
<td>0.9852</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td>SMC2(30)</td>
<td>0.9829</td>
<td>894</td>
</tr>
<tr>
<td></td>
<td>SMC3(30)</td>
<td>0.9894</td>
<td>949</td>
</tr>
<tr>
<td>3</td>
<td>SMC1(20)</td>
<td>0.9876</td>
<td>589</td>
</tr>
<tr>
<td></td>
<td>SMC2(20)</td>
<td>0.9863</td>
<td>661</td>
</tr>
<tr>
<td></td>
<td>SMC3(20)</td>
<td>0.9865</td>
<td>708</td>
</tr>
<tr>
<td></td>
<td>SMC1(30)</td>
<td>0.9896</td>
<td>643</td>
</tr>
<tr>
<td></td>
<td>SMC2(30)</td>
<td>0.9891</td>
<td>727</td>
</tr>
<tr>
<td></td>
<td>SMC3(30)</td>
<td>0.9905</td>
<td>785</td>
</tr>
</tbody>
</table>

The procedure SMC2 may eliminate inferior designs by multiple comparisons with a control or the two-sample-$t$ test, while SMC3 will eliminate inferior designs only by the two-sample-$t$ test. Note that SMC3 is no longer an indifference-zone R&S procedure, and it will allocate additional samples in order to ensure that the chosen design is the best. There are three settings in this experiment, settings 1, 2, and 3 corresponding to the set up in experiments 1, 2, and 3. Table 8 lists the results.

The sample-size upper bound of SMC3, computed from (12), has large variance and the procedure terminates with $r > N$ less than 20% of the time. If users are more concerned about $P(CS)$ than the runtime, the SMC3 procedure can be modified so that it terminates only when $|I| = 1$. On the other hand, SMC2 can eliminate inferior designs using Theorem 1, so it requires less data and has smaller observed $\hat{P}(CS)$’s, but still much greater than $P^*$. As expected, with smaller incremental sample sizes, these procedures require longer runtime, and obtain slightly smaller $\hat{P}(CS)$ with smaller sample sizes. This is because with a smaller incremental sample size, we increase both the number of iterations and the probability of eliminating inferior designs earlier. We believe the incremental sample sizes determined by
are a good balance between runtime and sample size. However, users can use different incremental samples depending on the computing cost to obtain sample. For example, if it takes a lot of computing resources to obtain samples, a smaller incremental size should be used to reduce the overall sample sizes.

5. CONCLUDING REMARKS

Many two-stage indifference-zone-selection procedures ignore a large amount of first-stage sampling information. The ETSS procedure utilizes both the means and variances from the first stage. The marginal computation effort required for the ETSS procedure is minimal, yet the achieved efficiency improvement is significant. Moreover, the ETSS procedure has an intrinsic subset pre-selection. Some drawbacks of two-stage R&S procedures are that we rely heavily on the information from the first stage, and the computation budget may not be enough to obtain all the required samples for the second stage.

We propose a simple sequential R&S procedure based on the ETSS procedure to avoid relying too much on sample-mean estimates from only one stage. The SRS procedure can minimize the sample size with a given $P^*$ or maximize $P(\text{CS})$ with a pre-determined computing budget. The incremental sample sizes of different designs at each iteration are computed dynamically according to the difference between sample means and sample variances at each iteration, and the indifference amount $d^*$. The SRS procedure allocates incremental sample sizes aggressively at earlier iterations, becomes less aggressive as the procedure proceeds, and effectively allocates additional replications or batches to more promising alternatives. Inferior designs are dropped from further consideration as the sequential procedure proceeds to reduce the overall computational effort. We improve the efficiency of the SRS procedure further by implementing all pairwise comparisons at each iteration.

The main advantage of SRS is that the algorithm determines the number of additional simulation replications based on both the means and variances, and approximately maintains the ratios required to maximize $P(\text{CS})$ with a given sample size at all iterations, so significantly improves the efficiency of R&S procedures. Unlike two-stage procedures, which are designed to guarantee $P(\text{CS})$ and assume all the required computing budget is available, the SRS procedure will be able to maximize $P(\text{CS})$ with a given computing budget.

Our experimental results show that the SRS procedure is a powerful tool for selecting the best $b'$ designs out of $k$ alternatives. The SRS procedure improves $P(\text{CS})$ with slightly larger total simulation replications or batches compared to the ETSS procedure. Furthermore,
incorporating all comparisons can improve the performance of R&S procedures in terms of 
P(CS) and sample size.

ACKNOWLEDGEMENTS

We would like to thank Dr. Chun-Hung Chen for our discussion of the approaches of OCBA 
and ETSS procedures, and the anonymous referees for their valuable comments.

APPENDIX A:

We evaluate P(CS) of the ETSS procedure assuming the true means are known. Without loss 
of generality, assume \( \mu_{i_1} + d^* \leq \mu_{i_2} \leq \ldots \leq \mu_{i_k} \). Let \( d_i = \max(d^*, \mu_i - \mu_{i_1}) \) and let the current 
sample size be \( r = n_0 \). Then

\[
P(CS) = P[X_{i_1}(N_{i_1}) < X_{i_l}(N_{i_l}), \text{ for } l = 2, 3, \ldots, k] \]
\[
= P\left[ \frac{X_{i_1}(N_{i_1}) - (X_{i_l}(N_{i_l}) - d_i)}{\sqrt{\sigma_{i_1}^2/N_{i_1} + \sigma_{i_l}^2/N_{i_l}}} < \frac{d_i}{\sqrt{\sigma_{i_1}^2/N_{i_1} + \sigma_{i_l}^2/N_{i_l}}}, \text{ for } l = 2, 3, \ldots, k \right].
\]

Denote

\[
Z_{i_l} = \frac{X_{i_1}(N_{i_1}) - (X_{i_l}(N_{i_l}) - d_i)}{\sqrt{\sigma_{i_1}^2/N_{i_1} + \sigma_{i_l}^2/N_{i_l}}}
\]

and

\[
Q_{i_l} = \frac{h}{\sqrt{\sigma_{i_l}^2/S_{i_l}^2(r) + \sigma_{i_1}^2/S_{i_1}^2(r)}}, l = 2, 3, \ldots, k.
\]

Since \( N_i \geq (h/d_i)^2S_{i_l}^2(r) \) and \( d_{i_1} \leq d_i \) for all \( i \),

\[
\frac{d_i}{\sqrt{\sigma_{i_1}^2/N_{i_1} + \sigma_{i_l}^2/N_{i_l}}} \geq \frac{d_i}{\sqrt{(d_i\sigma_{i_1})^2/(hS_{i_l}(r))^2 + (d_i\sigma_{i_l})^2/(hS_{i_1}(r))^2}}
\]
\[
\geq \frac{hd_i}{\sqrt{(d_i\sigma_{i_1})^2/S_{i_l}^2(r) + (d_i\sigma_{i_l})^2/S_{i_1}^2(r)}}
\]
\[
= Q_{i_l}.
\]

Thus,

\[
P(CS) \geq P[Z_{i_l} < Q_{i_l}, l = 2, 3, \ldots, k] \geq \prod_{l=2}^{k} P[Z_{i_l} < Q_{i_l}].
\]

The second inequality follows from Slepian’s inequality (Tong 1980) since the \( Z_i \)'s are positively 
correlated.
Let $\Phi(x)$ be the cdf (cumulative distribution function) of the standard normal distribution, let $g(x)$ be the pdf (probability density function) of a chi-square distribution with $r - 1$ df (degrees of freedom), and let $Q_{i_l} = Q$ for $l = 2, 3, \ldots, k$. We then follow the steps in the proof of Proposition 1 in Rinott (1978) to show that

$$
P(CS) \geq E(\Phi^{k-1}(Q)) = \int_0^\infty \Pi_{i=2}^k \left[ \int_0^\infty \Phi \left( \frac{h}{\sqrt{(r-1)(1/x + 1/y)}} \right) g(x) dx \right] g(y) dy. $$

The term on the right-hand side is the lower confidence bound (LCB) and is at its tightest under the LFC. As the controlled distance $d_i$ increase, the LCB becomes looser because the difference between $d_i$ and $d_{i_1}$ increases. This is a good property because as the $d_i$ increase we will have greater $P(CS)$ on the pairwise comparison between designs $i$ and $i_1$, for $i \neq i_1$.

**APPENDIX B:**

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

$$
P[Z^2 \leq \sum_{i=1}^k t_i^2 \gamma_i \frac{Y_i}{v_i}] \geq 1 - \alpha$$

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

Let $d_{ij} = \mu_i - \mu_j$, then

$$
P[\bar{X}_i > \bar{X}_j + w_{ij}] = 1 - P[\bar{X}_i \leq \bar{X}_j + w_{ij}]$$

$$= 1 - P\left[ \frac{(\bar{X}_i - \bar{X}_j) - d_{ij}}{\sqrt{\frac{\sigma_i^2}{r} + \frac{\sigma_j^2}{r}}} \leq \frac{w_{ij} - d_{ij}}{\sqrt{\frac{\sigma_i^2}{r} + \frac{\sigma_j^2}{r}}} \right]$$

$$\leq 1 - P\left[ \frac{(\bar{X}_i - \bar{X}_j) - d_{ij}}{\sqrt{\frac{\sigma_i^2}{r} + \frac{\sigma_j^2}{r}}} \leq \frac{v_{ij}}{\sqrt{\frac{\sigma_i^2}{r} + \frac{\sigma_j^2}{r}}} \right].$$

Denote

$$v_{ij} = \sqrt{\frac{\sigma_i^2}{r} + \frac{\sigma_j^2}{r}},$$

$$Z_{ij} = \frac{(\bar{X}_i - \bar{X}_j) - d_{ij}}{v_{ij}},$$

and

$$Q_{ij} = \frac{w_{ij}}{v_{ij}} = t_{1-\alpha,r-1} \sqrt{\frac{S_i^2(r) + S_j^2(r)}{\sigma_i^2 + \sigma_j^2}}.$$
Note that $Q_{ij} \geq 0$ and $Z_{ij}$ is $\mathcal{N}(0,1)$, so

$$\mathbb{P}[Z_{ij} \leq Q_{ij}] = \frac{1}{2} + \mathbb{P}[0 \leq Z_{ij} \leq Q_{ij}]$$

$$= \frac{1}{2}(1 + \mathbb{P}[Z_{ij}^2 \leq Q_{ij}^2])$$

$$= \frac{1}{2}(1 + \mathbb{P}[Z_{ij}^2 \leq t^2 \gamma_1 \frac{S_i^2(r)}{\sigma_i^2} + t^2 \gamma_2 \frac{S_j^2(r)}{\sigma_j^2}])$$

$$\geq 1 - \alpha$$

The inequality follows from the above Lemma with $\gamma_1 = 1 - \gamma_2 = \sigma_i^2/(\sigma_i^2 + \sigma_j^2)$. Therefore,

$$\mathbb{P}[^{\bar{X}}_i > \bar{X}_j + w_{ij}] \leq \alpha.$$

REFERENCES


