Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

PERFORMANCE EVALUATIONS OF COMPARISON-WITH-A-STANDARD PROCEDURES

E. Jack Chen

BASF Corporation
333 Mount Hope Avenue
Rockaway, New Jersey 07866, U.S.A.

ABSTRACT

We investigate the performance of a heuristic sequential procedure to compare a finite number of designs with respect to a single standard. The goal is to identify the best design, and if the chosen best design is not the standard, to determine whether the chosen best design is better than the standard. We give preferential status to the standard because there are costs and time involved in replacing the standard. We accomplish this goal by extending indifference-zone selection procedures. An experimental performance evaluation demonstrates the validity and efficiency of our sequential procedures.

1 INTRODUCTION

When evaluating alternative system designs, we are interested in selecting the best of a number of competing designs. In this paper, we consider comparison with a standard (control), i.e., one of the designs is designated as the standard, and the others are evaluated with respect to this standard. The goal is to identify the best design, and if the chosen best design is not the standard, to determine whether the chosen best design is better than the standard. In doing so, we would like to guarantee at least a lower bound on the probability of correct selection whenever the selected design satisfies the specified requirements. Let μ_0 denote the expected response of the standard design and μ_i denote the expected response of design i, for $i = 1, 2, \ldots, k$. The response of the standard design μ_0 may be known or unknown. In the statistics literature, a problem in which μ_0 is *known* is called a *comparison with a standard*, a problem in which μ_0 is the *unknown* mean of a control is referred to as a *comparison with a control*. Even though classical selection procedures cannot be applied directly to perform comparison with a standard, there has been extensive work done in this area, for instance, Nelson and Goldsman (2001) and Kim (2005). However, the procedure of Nelson and Goldsman (2001) requires pre-computed critical values, which often requires intensive numerical integration. Nelson and Goldsman (2001) suggest using a separate simulation experiment to *estimate* the critical values needed for their procedure to solve the problem at hand. Even though computer programs are available to estimate those critical values, the deviation of those critical values involves advanced mathematics. The procedure of Kim (2005) is effective in terms of sample size. However, it generally requires many iterations and thus long execution time because its incremental sample size is one.

Let μ_{i_l} be the l^{th} smallest of the μ_i's, so that $\mu_{i_1} \leq \mu_{i_2} \leq \cdots \leq \mu_{i_{k+1}}$. Our goal is to select a design with the smallest expected response μ_{i_1}. However, in practice if the difference between μ_{i_1} and μ_{i_2} is very small, we might not care if we mistakenly choose design i_2, whose expected response is μ_{i_2}. The "practically significant" difference d^* (a positive real number) between the best and a satisfactory design is called the *indifference zone* in statistical literature, and it represents the smallest difference that we care about. In a stochastic simulation, CS (correct selection) can never be guaranteed with certainty. The probability of CS denoted by P(CS), depends on sample sizes and becomes higher as sample sizes become larger. Parameter configurations satisfying $\mu_{i_2} - \mu_{i_1} \geq d^*$ are said to be in the *preference zone* for a correct selection; configurations satisfying $\mu_{i_2} - \mu_{i_1} < d^*$ are said to be in the indifference zone. Formally, we say a system i is *d^*-near-best* if μ_i is within a specified amount d^* of the smallest mean.

The indifference-zone approach wants to select a system i such that $\mu_i - \mu_{i_1} < d^*$, some literature refer to this event as the probability of good selection (P(GS)) and use P(CS) to indicate the event in which we select system i_1. In this paper, we do not distinguish the difference between the two and use P(CS) to indicate the event that we select a good design. In comparison with a standard, ideally we wish to retain the standard when $\mu_0 < \mu_{i_1} + d^*$ since there are costs and time involved in replacing the standard design with an alternative. However, large sample sizes are required to achieve correct pairwise comparisons when μ_0 and $\mu_{i_1} + d^*$

1-4244-0501-7/06/$20.00 ©2006 IEEE

are close together. In the spirit of the indifference-zone approach and the requirements that are discussed in Nelson and Goldsman (2001), we regard a correct comparison with a standard as: 1) we select design 0 when $\mu_0 \leq \mu_{i_1}$; 2) we select design i_1 when $\mu_{i_1} + d^* < \mu_i$ ($0 \leq i \leq k$, $i \neq i_1$); or 3) we select design i such that $\mu_i < \mu_{i_1} + d^*$ and $\mu_i < \mu_0$. Hence, for a d^*-near-best system to be regarded as a correct comparison with a standard, it must be better than the standard.

To obtain a pre-specified precision of the estimate for a design decision, a large number of samples (simulation replications) are often required for each design alternative. If the number of design alternatives is large, the total simulation run time will be significantly longer. Various schemes have been proposed to enhance the effectiveness of simulation experiments. Chen and Kelton (2005) show that if one or more very good alternatives are found early in the process, then they can be used to eliminate a greater number of inferior designs. In this paper, we present a variation of the comparison-with-a-standard procedure of Chen (2006).

The rest of this paper is organized as follows. In Section 2, we provide the background necessary for the proposed procedure. In Section 3, we present our methodology and proposed procedure for comparison with a standard. In Section 4, we list the procedure of Kim (2005). In Section 5, we give our empirical-experimental results. In Section 6, we make some concluding remarks.

2 BACKGROUND

In this section, we introduce the necessary notation and background:

X_{ij}: the independent and normally distributed observations 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,

n_i: the intermediate number of replications or batches for design i,

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

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

$\bar{X}_i$: the sample mean performance measure for design i shorthand for $\bar{X}_i(n_i)$,

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

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

2.1 Assessing P(CS)

In order to ensure the selection procedures obtain the pre-specified goal, we must be able to assess P(CS). In this section, we assess P(CS) with the *assumption* that the true means are known. Let $\phi(x)$ and $\Phi(x)$ denote the probability density and distribution function, respectively, of the standard normal distribution. Let $\delta_{i_l} = \mu_{i_l} - \mu_{i_1}$ for $l = 2, 3, \ldots, k$. Then

$$\begin{aligned} \mathrm{P(CS)} &= \mathrm{P}[\bar{X}_{i_1}(N_{i_1}) < \bar{X}_{i_l}(N_{i_l}),\ \text{for } l = 2,3,\ldots,k] \\ &= \mathrm{P}[\bar{X}_{i_1}(N_{i_1}) - \bar{X}_{i_l}(N_{i_l}) + \delta_{i_l} < \delta_{i_l}, \\ &\quad \text{for } l = 2,3,\ldots,k] \\ &\geq \Pi_{l=2}^{k} \mathrm{P}[\bar{X}_{i_1}(N_{i_1}) - \bar{X}_{i_l}(N_{i_l}) + \delta_{i_l} < \delta_{i_l}] \\ &= \Pi_{l=2}^{k} \Phi(\delta_{i_l}/\sqrt{\sigma_{i_l}^2/N_{i_l} + \sigma_{i_1}^2/N_{i_1}}). \end{aligned}$$

The inequality follows from Slepian's inequality (Tong 1980) since the values $\bar{X}_{i_1} - \bar{X}_{i_l}$ are positively correlated. The last equality follows from the fact that the variate

$$Z_{i_l} = \frac{\bar{X}_{i_1} - \bar{X}_{i_l} + \delta_{i_l}}{\sqrt{\sigma_{i_l}^2/N_{i_l} + \sigma_{i_1}^2/N_{i_1}}}$$

has a $N(0,1)$ distribution, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

The simultaneous one-tailed P^* c.i. (confidence interval) half-widths, with design i_1 as a control, are

$$w_{i_l} = z_h \sqrt{\sigma_{i_l}^2/N_{i_l} + \sigma_{i_1}^2/N_{i_1}},$$

where z_h is a critical value such that $\mathrm{E}(\Phi^{k-1}(z_h)) = P^*$. By the property of the c.i. half-width,

$$\mathrm{P}[\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) + w_{i_l} \geq \mu_{i_l} - \mu_{i_1}, \quad \text{for } l = 2,3,\ldots,k] \geq P^*.$$

To achieve

$$\mathrm{P}[\bar{X}_{i_l}(N_{i_l}) - \bar{X}_{i_1}(N_{i_1}) > 0,\ \text{for } l = 2,3,\ldots,k] \geq P^*,$$

the sample sizes N_i should be large enough so that $\mu_{i_l} - \mu_{i_1} > w_{i_l}$. Note that the half-width w_{i_l} depends on the sample sizes.

The sample sizes determined by traditional indifference-zone selection procedures achieve $w_{i_l} \leq d^*$ for $l = 2, 3, \ldots, k$. Without loss of generality, assume $\mu_{i_1} + d^* \leq \mu_{i_2} \leq \ldots \leq \mu_{i_k}$. Let $d_i = \max(d^*, \mu_i - \mu_{i_1})$ for designs $1 \leq i \leq k$. The indifference-zone selection procedures that take into account the difference of sample means attempt to achieve $w_{i_l} \leq d_{i_l}$ for $l = 2, 3, \ldots, k$. This sample size allocation rule establishes the *efficiency* aspect of the procedure, see Section 3.2.

Since the true means are unknown, Rinott (1978) derives the required sample sizes based on the LFC (least favorable configuration), i.e., $\mu_{i_1} + d^* = \mu_{i_2} = \ldots = \mu_{i_k}$. Consequently, Rinott's procedure is conservative, i.e., it often obtains higher than required P(CS) with larger than necessary sample sizes. Chen and Kelton (2005) approximate the difference of true means by the difference of sample means, which can significantly improve the efficiency of selection procedures. However, the procedure does not guarantee P(CS) $\geq P^*$. It is known that indifference-zone selection procedures also guarantee that the coverage of multiple comparisons with the best (MCB) c.i. with probability at least P^*; see Section 3.3.

2.2 Comparison-with-a-Standard Via All-Pairwise Comparisons

Standard indifference-zone selection procedures will ensure

$$\text{P[select design } i_1] \geq P^* \text{ whenever } \mu_{i_1} \leq \mu_i - d^*, \forall i \neq i_1. \qquad (1)$$

Comparison-with-a-standard procedures ensure

$$\text{P[select design 0]} \geq P^* \text{ whenever } \mu_0 \leq \mu_{i_1}; \qquad (2)$$

and (1). Our goals are achieved if Equations (1) and (2) hold individually.

The sample size allocation strategy is to ensure both the Type I (i.e., rejecting the null hypothesis when in fact it is true) and Type II (i.e., concluding that the null hypothesis is true when in fact it is false) errors are equal to or less than a specified amount; see Section 3.1. The allocated sample sizes are large enough so that the $P = 1-(1-P^*)/k$ c.i. half-width between designs 0 and $i \neq 0$ $w_{oi} \leq d^*/2$. Recall that the half-width w_{0i} depends on the variance of designs 0 and i and the sample sizes. Consequently, the procedure will eliminate the standard design only when $\bar{X}_0 > \bar{X}_i + w_{oi}$ and $w_{oi} \leq d^*/2$ (i.e., the procedure has concluded with high confidence that $\mu_0 > \mu_i$). Furthermore, if $w_{oi} \leq d^*/2$ and $\bar{X}_0 \leq \bar{X}_i + w_{oi}$, we will remove design i from further simulation (i.e., the procedure has concluded with high confidence that $\mu_0 \leq \mu_i$). The procedure, denoted as CAPC, is as follows.

Comparison-with-a-Standard Via All-Pairwise Comparisons CAPC:

1. Specify an initial sample size n_0, an indifference-zone parameter d^*, and a confidence level P^*.
2. Initialize the set I to include all $k+1$ designs. Simulate n_0 replications or batches for each design $i \in I$. Set the iteration number $l = 0$, and $r = N_{0,l} = N_{1,l} = \cdots = N_{k,l} = n_0$, where $N_{i,l}$ is the sample size allocated for design i at the l^{th} iteration. Set $P = 1-(1-P^*)/k$.
3. Perform all pairwise comparisons and remove inferior design $j \neq 0$, i.e., $\bar{X}_j > \bar{X}_i + w_{ij}$ for some $i \in I$, from I. Remove the standard design 0 when $\bar{X}_0 > \bar{X}_i + w_{i0} + d^*/2$; or $\bar{X}_0 > \bar{X}_i + w_{i0}$ and $w_{i0} \leq d^*/2$. Here w_{ij} is the one-tailed P c.i. half-width.
4. If $w_{ij} < d^*$ and $\bar{X}_j > \bar{X}_i$, remove design $j \neq 0$ from I. If $w_{i0} \leq d^*/2$ and $\bar{X}_0 \leq \bar{X}_i + w_{i0}$ for some $i \in I$, remove design i from I.
5. If there is only one element (or the pre-determined number of best designs) in I, go to step 9.
6. Compute the critical value $h_t = \sqrt{2} t_{P,r-1}$.
7. Let $\bar{X}_{b,l} = \min_{i \in I} \bar{X}_{i,l}$ and let $U(\bar{X}_{b,l})$ denote the upper one-tailed P^* confidence limit of μ_b at the l^{th} iteration. If $|I| = 2$ and $0 \in I$, for all $i \in I$, set $\hat{d}_{i,l} = \max(d^*/2, \bar{X}_{i,l} - U(\bar{X}_{b,l}))$. Otherwise, set $\hat{d}_{i,l} = \max(d^*, \bar{X}_{i,l} - U(\bar{X}_{b,l}))$. Compute
$$\delta_{i,l+1} = \lceil ((h_t S_i(r)/\hat{d}_{i,l})^2 - r)^+ \rceil.$$
Here $(x)^+$ denotes $\max(0,x)$.
8. Set $l = l+1$. If $\delta_{i,l} = 0$, set $\delta_{i,l} = 1$. Set the incremental sample size at the l^{th} iteration $\delta_l = \min_{i \in I} \delta_{i,l} + 1$ and set $r = r + \delta_l$. For $\forall i \in I$, simulate additional δ_l samples, set $N_{i,l} = r$. Go to step 3.
9. If the standard design is in I, return the values 0 and $\bar{X}_0(N_0)$. Otherwise, return the values b and $\bar{X}_b(N_b)$, where $\bar{X}_b(N_b) = \min \bar{X}_i(N_i)$, $1 \leq i \leq k$ and i was not eliminated by all pairwise comparisons.

If $\mu_0 \leq \mu_i$ for $1 \leq i \leq k$ and d^* is significant large, there is only $(1-P^*)/k$ probability that the standard design is eliminated by mistake when compared with some alternative design i. Consequently, P[design 0 is selected] $\geq P^*$. On the other hand, if $\mu_{i_1} + d^* \leq \mu_i$ for all $i \neq i_1$, then P[design i is selected] $\leq (1-P^*)/k$ for some $i \neq i_1$. Note that the CAPC procedure also guarantees that the selected design b having $\mu_b < \mu_{i_1} + d^*$ with high confidence (i.e., the event $\mu_b < \mu_{i_1} + d^*$ is true $100P^*\%$ of the times).

If the sample sizes are large enough so that the one-tailed P CI half-width $w_{0i} \leq d^*/2$, then the precision of pairwise comparison between designs 0 and i can be guaranteed; see Section 3.1. To avoid allocating larger than necessary sample sizes, the procedure sets $\hat{d}_{i,l} = \max(d^*/2, \bar{X}_{i,l} - U(\bar{X}_{b,l}))$ when $|I| = 2$ and $0 \in I$. If $\hat{d}_{i,l} \leq d^*/2$, the allocated sample sizes should be large enough to achieve $w_{0i} \leq d^*/2$. The procedure then removes the standard design from further simulation when $\bar{X}_0 > \bar{X}_i + w_{0i}$; otherwise, the procedure removes the competing alternative. However, if $\hat{d}_{i,l} > d^*/2$, the allocated sample sizes probably are not large enough to achieve $w_{0i} \leq d^*/2$. To reduce the number of iterations without increasing the probability of eliminating the standard design by mistake, if $w_{oi} > d^*/2$, the procedure removes the standard design from further simulation only when $\bar{X}_0 > \bar{X}_i + w_{0i} + d^*/2$. If $\mu_0 \leq \mu_i$, then $\mathrm{P}[\bar{X}_0 > \bar{X}_i + w_{0i}] \leq 1 - P$. However, if the number of iterations is large, the event $\bar{X}_0 > \bar{X}_i + w_{0i}$ is likely to occur. That is, the probability of committing a Type I error increases as the number of comparison increases. The amount $d^*/2$ is added to reduce the chance of wrongly eliminating the standard (i.e., committing a Type I error).

3 METHODOLOGIES

Like most ranking and selection procedures, the proposed comparison-with-a-standard procedure also requires the input data to be independent and identically distributed (i.i.d.) normal. However, the variance can be different across designs. Many performance measures of interest are taken over some average of a sample path or a batch of samples. Thus, many applications tend to have a normally distributed simulation output. Users can use batch means (see Law and Kelton 2000) to obtain samples that are essentially i.i.d. normal if the nonnormality of the samples is a concern.

3.1 Two-Sample Tests

The conventional statistic for determining the significance of a difference of means is by null hypothesis test. It is known that the random variable $Y_i = \bar{X}_i(N_i) - \bar{X}_b(N_b)$ $(i \neq b)$ has approximately a t distribution with f_i :

$$\frac{(S_i^2(N_i)/N_i + S_b^2(N_b)/N_b)^2}{(S_i^2(N_i)/N_i)^2/(N_i-1) + (S_b^2(N_b)/N_b)^2/(N_b-1)} \tag{3}$$

d.f. (degrees of freedom); see Law and Kelton (2000) for detail. Furthermore, if $N_i = N_b = r$, it is fairly safe to approximate the value of $\bar{X}_i(r) - \bar{X}_b(r)$ with a t distribution with $r-1$ d.f.; see Scheffé (1970) for further discussion. Chen (2004) derives an indifference-zone selection procedure based on this principle.

Suppose the sample sizes are n_0 and n_b for designs 0 and b, respectively. The test at confidence level $1-\alpha$ of $H_0 : \mu_0 \leq \mu_b$ against the alternative $H_1 : \mu_0 > \mu_b$ is based on the test statistic

$$T = \frac{\bar{X}_0 - \bar{X}_b}{\sqrt{S_0^2/n_0 + S_b^2/n_b}}.$$

The acceptance region for this test is $T \leq t_{1-\alpha,f}$, where $t_{1-\alpha,f}$ is the $1-\alpha$ quantile of the t distribution with f d.f. If $\mu_0 - \mu_b \geq d^*$, the probability of committing a Type II error β is

$$\mathrm{P}[\bar{X}_0 - \bar{X}_b \leq t_{1-\alpha,f}\sqrt{S_0^2/n_0 + S_b^2/n_b}].$$

We have

$$\mathrm{P}[\frac{(\bar{X}_0 - \bar{X}_b) - d^*}{\sqrt{S_0^2/n_0 + S_b^2/n_b}} \leq \frac{t_{1-\alpha,f}\sqrt{S_0^2/n_0 + S_b^2/n_b} - d^*}{\sqrt{S_0^2/n_0 + S_b^2/n_b}}]$$
$$\leq F(t_{1-\alpha,f} - \frac{d^*}{\sqrt{S_0^2/n_0 + S_b^2/n_b}}),$$

where F is the cdf (cumulative distribution function) of the t distribution with f d.f. Thus, the probability that the test statistic falls in the acceptance region is

$$\beta \leq F(t_{1-\alpha,f} - \frac{d^*}{\sqrt{S_0^2/n_0 + S_b^2/n_b}}).$$

For fixed d^* and α, β can be evaluated as a function of sample sizes n_0 and n_b. For more detail, see Rice (1995). Suppose we want to limit the probability of β, the sample sizes n_0 and n_b should be large enough such that

$$t_{1-\alpha,f} - \frac{d^*}{\sqrt{S_0^2/n_0 + \sigma_b^2/n_b}} = t_{\beta,f}.$$

If we choose $\alpha = \beta < 0.5$, then

$$2t_{1-\alpha,f} = \frac{d^*}{\sqrt{S_0^2/n_0 + \sigma_b^2/n_b}}.$$

Hence, the sample sizes should be large enough such that the one-tailed $1-\alpha$ c.i. half-width

$$w = t_{1-\alpha,f}\sqrt{S_0^2/n_0 + \sigma_b^2/n_b} = d^*/2.$$

3.2 Improving the Efficiency

Branke et al. (2005) evaluate ranking and selection procedures with the following aspects:

- Efficiency: The mean evidence for correct selection as a function of the mean number of samples.
- Controllability: The ease of setting a procedure's parameters to achieve a targeted evidence level (as opposed to a potentially conservative guarantee that the targeted evidence level is exceeded).

- Robustness: The dependency of a procedure's effectiveness on the underlying problem characteristics.
- Sensitivity: The effect of the parameters on the mean number of samples needed.

Another performance measure of selection procedures is the *execution time*, especially the runtime of ranking and selection (as opposed to the runtime of generating samples). Many samples can be generated simultaneously when deploying ranking and selection in a parallel and distributed environment; see Chen (2005). Furthermore, the runtime of ranking and selection is correlated with the number of iterations. For sequential procedures to work efficiently, a good incremental sample size must be used. With a small incremental sample size, the procedure needs to iterate the computation steps many times. On the other hand, with a large incremental sample size, we are putting too much confidence on the mean and variance estimators of early iterations and can result in waste of computation time to obtain an unnecessarily high confidence level of non-critical designs.

We develop a new strategy to calculate the incremental size dynamically at each iteration to further improve the efficiency of the procedure. The procedure in Section 2.2 uses the same incremental sample size for each design at each iteration to reduce computation effort of performing pairwise comparisons and computing the critical constant h_t. However, the number of iterations may be greater than desired in certain situations; for example, when the true means are monotone increasing or decreasing. We propose to compute the incremental sample size for each design by

$$\delta_{i,l+1} = \lceil ((h_t S_i(N_{i,l})/\hat{d}_{i,l})^2 - N_{i,l})^+/2 \rceil.$$

Thus, the sequential procedure allocates incremental sample sizes aggressively at earlier iterations and become less aggressive as the procedure proceeds and brings us closer to the optimal solution. This way we will be able to reduce the number of iterations without the risk of putting too much resources to simulate non-critical designs. Chen and Kelton (2005) show that this incremental sample size allocation strategy has good properties, i.e., the ratios of allocated samples are close to that of Optimal Computing Budget Allocation of Chen et al. (2000).

With this sample size increment strategy, the allocated sample sizes for the surviving designs are likely to be different. Consequently, we need to compute the d.f. f by (3) for each pairwise comparison. Furthermore, the critical value $h_t = \sqrt{2}t_{P,r-1}$ will be conservatively computed with $r = \min_{i \in I} N_{i,l}$ at iteration l. The modified procedure is denoted CAPC2 in the remainder of the paper.

3.3 Inference from the Comparison-with-a-Standard Procedure

Multiple comparisons provide simultaneous confidence intervals on selected differences among the designs. It is known that indifference-zone selection procedures also guarantee that the coverage of MCB c.i. with probability at least P^*. These c.i.'s bound the differences between the performance of each design and the best of the others with a prespecified confidence level.

Since there are $k+1$ designs under consideration, in the following discussion w_{ij} is the one-tailed $P = 1-(1-P^*)/k$ c.i. half-width. The multiple comparison with the best confidence intervals are

$$\begin{aligned}&\mathrm{P}[\mu_i - \min_{j\neq i}\mu_j \in \\ &[\max_{j\neq i}(\bar{X}_i - \bar{X}_j - w_{ij})^-, \max_{j\neq i}(\bar{X}_i - \bar{X}_j + w_{ij})^+], \forall i] \geq P^*.\end{aligned}$$

Here $(x)^-$ denotes $\min(0,x)$. We follow the discussion of Nakayama (1997) to construct MCB intervals. Define the events

$$E = \{\mu_i - \mu_{i_1} \leq \bar{X}_i - \bar{X}_{i_1} + w_{ii_1}, \forall i \neq i_1\},$$

$$E_L = \{\mu_i - \min_{j\neq i}\mu_j \geq \max_{j\neq i}(\bar{X}_i - \bar{X}_j - w_{ij})^-, \forall i\},$$

$$E_U = \{\mu_i - \min_{j\neq i}\mu_j \leq \max_{j\neq i}(\bar{X}_i - \bar{X}_j + w_{ij})^+, \forall i\},$$

$$\begin{aligned}E_T = \{&\mu_i - \min_{j\neq i}\mu_j \in \\ &[\max_{j\neq i}(\bar{X}_i - \bar{X}_j - w_{ij})^-, \max_{j\neq i}(\bar{X}_i - \bar{X}_j + w_{ij})^+], \forall i\}.\end{aligned}$$

Note that E is the event that the upper one-tailed confidence intervals for Multiple Comparisons with a control, with the control being design i_1, contain all of the true differences $\mu_i - \mu_{i_1}$. Since $\mathrm{P}[\mu_i - \mu_{i_1} \leq \bar{X}_i - \bar{X}_{i_1} + w_{ii_1}] \geq P$ $\forall i$, $\mathrm{P}[E] \geq P^*$. Now, following an argument developed by Edwards and Hsu (1983), we have that $E \subset E_L \cap E_U$, which will establish the result $\mathrm{P}[E_T] \geq P^*$. First we prove that $E \subset E_L$:

$$\begin{aligned}E &\subset \{\mu_{i_1} - \mu_j \geq \bar{X}_{i_1} - \bar{X}_j - w_{i_1 j}, \forall j \neq i_1\} \\ &\subset \{\mu_{i_1} - \mu_{i_2} \geq \bar{X}_{i_1} - \bar{X}_j - w_{i_1 j}, \forall j \neq i_1\} \\ &\subset \{\mu_i - \mu_{i_2} \geq \max_{j\neq i}(\bar{X}_i - \bar{X}_j - w_{ij})^-, \forall i\} \\ &\subset \{\mu_i - \min_{j\neq i}\mu_j \geq \max_{j\neq i}(\bar{X}_i - \bar{X}_j - w_{ij})^-, \forall i\},\end{aligned}$$

where the second step follows since $\mu_{i_1} - \mu_{i_2} \geq \mu_{i_1} - \mu_j$ for all $j \neq i_1$ and the third step follows since $\mu_i - \mu_{i_2} \geq 0$

for all $i \neq i_1$ and $(x)^- \leq 0$. Now we show $E \subset E_U$.

$$\begin{aligned} E &\subset \{\mu_i - \mu_{i_1} \leq \max_{j \neq i}(\bar{X}_i - \bar{X}_j + w_{ij}), \forall i \neq i_1\} \\ &\subset \{\mu_i - \min_{j \neq i} \mu_j \leq \max_{j \neq i}(\bar{X}_i - \bar{X}_j + w_{ij})^+, \forall i\}, \end{aligned}$$

where the first step follows since $\max_{j \neq i}(\bar{X}_i - \bar{X}_j + w_{ij}) \geq \bar{X}_i - \bar{X}_{i_1} + w_{ii_1}$ for all $i \neq i_1$ and the last step follows since $\mu_{i_1} - \min_{j \neq i_1} \mu_j \leq 0$ and $(x)^+ \geq 0$. Hence, $E \subset E_L \cap E_U$, and the proof is complete. See Edwards and Hsu (1983), Nelson and Matejcik (1995), and Nakayama (1997) for more details on multiple comparisons.

Traditional indifference-zone selection procedures achieve $w_{ij} \leq d^*$ and the MCB c.i. is simplified to

$$\begin{aligned} E_T = \{&\mu_i - \min_{j \neq i} \mu_j \in \\ &[(\bar{X}_i - \min_{j \neq i} \bar{X}_j - d^*)^-, (\bar{X}_i - \min_{j \neq i} \bar{X}_j + d^*)^+], \forall i\}. \end{aligned}$$

However, these tight c.i.'s come at a cost. Our procedure takes into account the differences of sample means, hence, the c.i. half-width w_{ii_1} is around $\max(d^*, \mu_i - \mu_{i_1})$ instead of d^*.

4 Fully Sequential Procedure of Kim (2005)

We compare the performance of our procedures with the FSP procedure of Kim (2005) in our empirical experiments. The FSP procedure is as follows.

1. **Setup**: Based on the input parameters: confidence level P^*, indifference-zone parameter d^* and first-stage sample size $n_0 \geq 2$. Calculate η and c as described below in **Constants**.
2. **Initialization**: Let $I = \{0, 1, 2, \ldots, k\}$ be the set of designs still in contention. Obtain n_0 observations X_{ij}, $j = 1, 2, \ldots, n_0$, from each design $i = 0, 1, 2, \ldots, k$. For all $i \neq l$, $i, l = 0, 1, 2, \cdots, k$ compute S_{il}^2, the sample variance of the difference between design i and design l, and let

$$a_{il} = \frac{\eta(n_0 - 1)S_{il}^2}{D_{il}} \text{ and } \lambda_{il} = \frac{D_{il}}{2c}$$

where

$$D_{il} = \begin{cases} d^*/2, & \text{if } i = 0 \text{ or } l = 0 \\ d^*, & \text{otherwise.} \end{cases}$$

3. **Screening**: For each $i \neq l$, $i \in I$, and $l \in I$,

$$\text{if } \sum_{j=1}^{r} (\mathcal{X}_{ij} - \mathcal{X}_{lj}) < \max\{0, -a_{il} + \lambda_{il} r\},$$

then eliminate i from I, where

$$\mathcal{X}_{qj} = \begin{cases} X_{qj} + d^*/2, & \text{if } q = 0 \\ X_{qj}, & \text{otherwise.} \end{cases}$$

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 $r = r + 1$.
5. **Constants**: The constant c may be any nonnegative integer. The constant η is the solution to the equation

$$\begin{aligned} &\sum_{l=1}^{c} (-1)^{l+1}(1 - \frac{1}{2}\mathcal{I}(l = c)) \times \\ &\qquad (1 + \frac{2\eta(2c - l)l}{c})^{\frac{-(n_0-1)}{2}} = \beta, \quad (4) \end{aligned}$$

where $\mathcal{I}$ is the indicator function and β is selected so that the overall confidence is P^*. When designs are simulated independently, the procedure sets $\beta = 1 - (P^*)^{1/k}$. When common random numbers (CRN) are used, the procedure sets $\beta = (1 - P^*)/k$.

Both FSP and CAPC procedures perform all-pairwise comparisons to eliminate inferior designs early in the iterations. However, the FSP procedure eliminates inferior designs based on whether the partial sum is within the continuation region and is valid regardless of the indifference amount. On the other hand, the CAPC procedure eliminates inferior designs based on the two-sample-t tests and achieves the statistical guarantee only when the indifference amount d^* is reasonably significant, say larger than 10% of the standard error of the difference between the performance measures.

The FSP procedure is efficient in terms of sample sizes. However, a long runtime is required because the incremental sample size is one. The CAPC procedure determines the incremental sample sizes dynamically based on the underlying designs and are very effective in terms of the number of iterations.

5 EMPIRICAL EXPERIMENTS

In this section, we present some empirical results. 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. In order to compare our comparison-with-a-standard procedures with other known procedures, we use similar experimental designs of Kim (2005). We chose the first-stage sample size to be $n_0 = 10$.

The number of designs under consideration is $k = 5$. The indifference zone, d^*, was set to $d^* = 1/\sqrt{n_0}$ and we set the variance of the best design (either design 0 or 1) to one. In this setting, d^* is the standard deviation of the first-stage sample mean of the best design. The minimal P(CS) of P^* is set to 0.95. The LFC, equal means configuration (EMC), and the monotonic increasing means (MIM) configurations were used; see Table 1 for the corresponding values. The variance of non-best designs are either monotonic increasing or monotonic decreasing; see Table 2 for the configuration. The minimum P(CS) should occur at the LFC and EMC configurations. In the EMC, a CS means retaining the standard, while in the LFC it means selecting design 1. The MIM configuration is to demonstrate the effectiveness of the procedures in eliminating inferior designs, whose $\bar{X}_i$ is far in excess of $\bar{X}_b$, at early iterations.

5.1 Experiments of Comparisons with a Control

In this experiment, the true mean of the standard μ_0 is unknown and needs to be simulated with alternative designs. The results are based on 10000 independent simulation runs. For comparison, we include the average sample sizes allocated by NG (Nelson and Goldsman 2001) and FSP (Kim 2005).

Table 3 lists the results of experiment 1 where design 0 is the best. The NG, FSP, CAPC, and CAPC2 rows list the average sample size of the corresponding procedures. The Iteration row lists the average number of iterations. The Stdev row lists the standard deviation of the average sample sizes. The $\hat{P}$(CS) row lists the proportion of correction selection of these 10000 simulation runs. Even though the $\hat{P}$(CS)'s of NG and FSP are not listed here, all $\hat{P}$(CS)'s are greater than the specified nominal level of 0.95. Procedure NG is conservative, thus, it allocates large sample sizes and achieves high P(CS). Furthermore, NG is developed based on the LFC, thus, it allocates roughly the same samples under the MIM and the EMC (LFC) configurations. All other procedures take into account the information of sample means and allocate less samples under the MIM configuration. Even though CAPC and CAPC2 generally allocates more samples than FSP, the number of iterations is significantly smaller.

The average number of iterations of CAPC and CAPC2 under EMC with increasing variance is 63 and 11, respectively. On the other hand, the average number of iterations of FSP under the same setting will be greater than 364 ($\approx 2239/6-10$). Therefore, FSP generally requires much longer execution time than CAPC and CAPC2. CAPC and CAPC2 determine the incremental sample size dynamically and are very efficient in terms of the number of iterations.

Table 4 lists the results of experiment 2 where design 1 is the best. The observed $\hat{P}(CS)$ of CAPC under the LFC with increasing variance is 0.9469 and the observed $\hat{P}(CS)$ of CAPC2 under the LFC with increasing and decreasing variance are 0.9465 and 0.9493, respectively. Even though these $\hat{P}(CS)$ are below the nominal value of 0.95, they are close to the nominal value. We believe this is because of the stochastic nature of the experiment and an indication of the controllability of these procedures. All other $\hat{P}$(CS)'s are greater than the nominal value. Again, CAPC and CAPC2 requires slightly larger sample sizes to achieve the required precision when compared with FSP. CAPC2 generally requires smaller sample sizes and smaller number of iterations when compared with CAPC.

5.2 Experiments of Comparisons with a Standard

In this experiment, we assume the true mean of the standard μ_0 is known and does not need to be simulated with alternative designs. Consequently, the variance of the standard $\sigma_0^2 = 0$. In every other respect the experiments were conducted as described above. Tables 5 and 6 list the experimental results. In general all procedures allocate less samples when the true mean of the standard is known since they don't allocate any samples for the standard. These results are generally similar to those experiments when the true mean of the standard is unknown. The observed $\hat{P}(CS)$ of CAPC and CAPC2 under the EMC with increasing variance respectively are 0.9482 and 0.9490 just below the nominal value of 0.95. All other $\hat{P}(CS)$ are higher than the nominal value. In the EMC setting, the standard design is incorrectly eliminated at the initial iteration most of the times, hence, we recommend using a larger initial sample size, for example $n_0 \geq 20$. When design 0 is the best, CAPC allocates less samples and achieves lower $\hat{P}(CS)$ when compared with the μ_0 is unknown cases. On the other hand, when design 1 is the best CAPC and CAPC2 generally achieve higher $\hat{P}(CS)$ with smaller sample sizes when compared with the μ_0 is unknown cases. In general, CAPC2 allocates smaller sample sizes and requires a smaller number of iterations than CAPC.

6 CONCLUSIONS

We have presented a sequential procedure for comparison with a standard based on the procedure of Chen (2006). Our procedures allow for unequal variances across designs and known or unknown expected performance of the standard.

The procedure incorporates all pairwise comparisons to eliminate inferior designs at each iteration, which may reduce the overall computational effort. The procedure is robust to minor departure of the normality assumption. Furthermore, these procedures are derived based on the Bonferroni inequality, so one can use common random numbers to increase the P(CS) without any further assumptions. However, Nelson and Goldsman (2001) point out that it may be counterproductive to use CRN to perform comparison with

Table 1: Mean Configuration

Best	Model	μ_0	μ_1	μ_2	μ_3	μ_4	μ_5
Design 0	EMC	0	0	0	0	0	0
	MIM	$-d^*$	0	d^*	$2d^*$	$3d^*$	$4d^*$
Design 1	LFC	0	$-d^*$	0	0	0	0
	MIM	0	$-d^*$	0	d^*	$2d^*$	$3d^*$

Table 2: Variance Configuration

Best	Model	μ_0	μ_1	μ_2	μ_3	μ_4	μ_5
Design	Increase	1	$1+d^*$	$1+2d^*$	$1+3d^*$	$1+4d^*$	$1+5d^*$
0	Decrease	1	$\frac{1}{1+d^*}$	$\frac{1}{1+2d^*}$	$\frac{1}{1+3d^*}$	$\frac{1}{1+4d^*}$	$\frac{1}{1+5d^*}$
Design	Increase	$1+d^*$	1	$1+d^*$	$1+2d^*$	$1+3d^*$	$1+4d^*$
1	Decrease	$\frac{1}{1+d^*}$	1	$\frac{1}{1+d^*}$	$\frac{1}{1+2d^*}$	$\frac{1}{1+3d^*}$	$\frac{1}{1+4d^*}$

Table 3: P(CS) and Sample Sizes When Design 0 is the Best

Procedure	MIM	MIM	EMC	EMC
Type	Increas	Decreas	Increas	Decreas
NG	4615	1623	4618	1603
FSP	599	360	2239	867
CAPC	777	563	1835	884
CAPC2	648	388	1689	774
CAPC				
Iteration	4	4	63	26
Stdev	380	288	313	169
$\hat{P}(CS)$	0.9998	0.9995	0.9558	0.9672
CAPC2				
Iteration	3	3	11	9
Stdev	268	169	408	191
$\hat{P}(CS)$	0.9992	0.9995	0.9509	0.9581

Table 4: P(CS) and Sample Sizes When Design 1 is the Best

Procedure	MIM	MIM	LFC	LFC
Type	Increas	Decreas	Increas	Decreas
NG	4032	1773	4056	1747
FSP	991	711	1225	780
CAPC	1065	741	1263	821
CAPC2	1107	718	1383	778
CAPC				
Iteration	22	20	38	25
Stdev	364	263	386	271
$\hat{P}(CS)$	0.9737	0.9791	0.9469	0.9650
CAPC2				
Iteration	26	19	26	20
Stdev	362	240	380	248
$\hat{P}(CS)$	0.9709	0.9730	0.9465	0.9493

Table 5: P(CS) and Sample Sizes When Design 0 is the Best and μ_0 is Known

Procedure	MIM	MIM	EMC	EMC
Type	Increas	Decreas	Increas	Decreas
NG	2279	647	2291	644
FSP	322	124	1283	365
CAPC	507	422	1278	545
CAPC2	456	286	1037	438
CAPC				
Iteration	5	4	16	10
Stdev	215	199	309	141
$\hat{P}(CS)$	0.9869	0.9874	0.9482	0.9567
CAPC				
Iteration	3	3	7	6
Stdev	177	108	182	78
$\hat{P}(CS)$	0.9864	0.9866	0.9490	0.9502

Table 6: P(CS) and Sample Sizes When Design 1 is the Best and μ_0 is Known

Procedure	MIM	MIM	LFC	LFC
Type	Increas	Decreas	Increas	Decreas
NG	1918	802	1932	795
FSP	419	307	673	384
CAPC	638	516	834	585
CAPC2	605	371	870	440
CAPC				
Iteration	7	6	18	12
Stdev	245	194	247	204
$\hat{P}(CS)$	0.9875	0.9882	0.9646	0.9704
CAPC2				
Iteration	5	4	6	5
Stdev	204	112	194	113
$\hat{P}(CS)$	0.9828	0.9817	0.9591	0.9599

a standard when μ_0 is known and is not simulated with alternatives. Moreover, this procedure can be deployed in a parallel and distributed environment to shorten the duration of execution time, see Chen (2005).

We have shown that the proposed procedures are versatile and easy to apply. Our approach is easy to state, interpret, and implement. These procedures preserve the simple structure of indifference-zone selection while being more efficient in situations where there are many alternative designs but some are not really competitive.

ACKNOWLEDGMENTS

The author thanks Bobbie Chern for helpful comments.

REFERENCES

Branke, J., S. Chick, and C. Schmidt. 2005. New developments in ranking and selection: an empirical comparison of the three main approaches. In *Proceedings of the 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 708-717.

Chen, E. J. 2004. Using ordinal optimization approach to improve efficency of selection procedures. *Journal of Discrete Event Dynamic Systems* 14(2):153-170.

Chen, E. J. 2005. Using Parallel and Distributed Computing to Increase the Capability of Selection Procedures. In *Proceedings of the 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 723-731.

Chen, E. J. 2006. Comparison with a standard via all-pairwise comparisons. *Journal of Discrete Event Dynamic Systems* 16(3):385-403.

Chen, E. J., and W. D. Kelton. 2005. Sequential selection procedures: using sample means to improve efficiency. *European Journal of Operational Research* 166(1):133-153.

Chen, C. H., J. Lin, E. Yücesan, and S. E. Chick. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Journal of Discrete Event Dynamic Systems* 10: 251–270.

Edwards, D. G., and J. C. Hsu. 1983. Multiple comparisons with the best treatment. *Journal of the American Statistical Association* 78:965–971.

Kim, S.-H. 2005. Comparison with a standard via fully sequential procedures. *ACM Transactions on Modeling and Computer Simulation* 11:251-273.

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

Nakayama, M. K. 1997. Multiple-comparison procedures for steady-state simulations. *Annals of Statistics* 25: 2433–2450.

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

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

Rice, J. A. 1995. *Mathematical statistics and data analysis*. 2nd ed. Belmont, California: Duxbury Press.

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

Scheffé, H. 1970. Practical solutions of the Behrens-Fisher problem, *Journal of the American Statistical Association* 65:1501–1508.

Tong, Y. L. 1980. *Probability inequalities in multivariate distributions.* New York:Academic Press.

AUTHOR BIOGRAPHY

E. JACK CHEN is a Senior Staff Specialist with BASF Corporation. He received a Ph.D. from the University of Cincinnati. His research interests are in the area of computer simulation. His email address is <e.jack.chen@basf.com>.