OPERATIONS RESEARCH
Vol. 54, No. 4, July–August 2006, pp. 743–755
ISSN 0030-364X | EISSN 1526-5463 | 06 | 5404 | 0743

DOI 10.1287/opre.1060.0311

Controlled Sequential Bifurcation: A New Factor-Screening Method for Discrete-Event Simulation

Hong Wan
School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907-2023, hwan@purdue.edu

Bruce E. Ankenman, Barry L. Nelson
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208-3119
{ankenman@northwestern.edu, nelsonb@northwestern.edu}

Screening experiments are performed to eliminate unimportant factors so that the remaining important factors can be more thoroughly studied in later experiments. Sequential bifurcation (SB) is a recent screening method that is well suited for simulation experiments; the challenge is to prove the "correctness" of the results. This paper proposes controlled sequential bifurcation, a procedure that incorporates a hypothesis-testing approach into SB to control error and power. A detailed algorithm is given, conditions that guarantee performance are provided, and an empirical evaluation is presented.

Subject classifications: simulation: design of experiments, statistical analysis, factor screening, sequential bifurcation.
Area of review: Simulation.
History: Received March 2004; revisions received October 2004, May 2005, June 2005; accepted July 2005.

1. Introduction

Screening experiments are designed to investigate the controllable factors in an experiment with a view toward eliminating the unimportant ones. According to the sparsity of effects principle, in many cases only a few factors are responsible for most of the response variation (Myers and Montgomery 2002). A good screening procedure should correctly and efficiently identify important factors. This is especially useful when the system is complicated and many factors are being considered.

In this paper, we focus on factor-screening methods for discrete-event simulations. Simulation experiments are different from physical experiments in at least three ways:

1. Screening problems in simulation can involve many more factors than real-world problems. In typical physical experiments it is difficult to control more than 15 factors, while in simulation experiments it is easy to control and simulate many decision variables because the experiment can be automated (Bettonvil and Kleijnen 1997; Kleijnen et al. 2005; Trocine and Malone 2000, 2001; Morris 2006).
2. In physical experiments, switching from one factor setting to another can be costly (time and money). In simulation, however, switching decision variable values is comparatively easy. This makes sequential methods especially attractive in simulation.
3. In simulation experiments, common random numbers (CRN) can be implemented to reduce the variance of estimated effects as compared to independent simulations (Law and Kelton 2000). Controlling random number seeds is not applicable in physical experiments, although the concept is similar to "blocking."

These differences suggest that screening strategies for simulation experiments will be different from those for physical experiments.

Many screening strategies have been developed to identify important factors with an economical number of design points and replications (Trocine and Malone 2000, 2001; Morris 2005). For instance, the first stage of response surface methodology is usually factor screening, which is often based on a first-order design, such as a Plackett-Burman design. There has been considerable research in this area (e.g., Myers and Montgomery 2002, Wu and Hamada 2000). However, most of these experiment-design strategies emphasize physical experiments and do not take advantage of the highly sequential nature of simulation experiments. In fact, recent research has gone in the opposite direction by combining the screening experiments and a follow-up response exploration into one design to screen out the important factors and build the model simultaneously (Cheng and Wu 2001).

Group-screening methods have been widely used for situations with large numbers of factors. The fundamental idea is to identify the important/unimportant factors as a group to save experimental effort (Lewis and Dean 2001). If a group is considered to be important, then subgroups or individual factors within the group should be further screened; if a group is not considered to be important, then the whole group can be classified as unimportant. In group screening, the effects of the factors that are grouped together must have the same sign to avoid cancellation, and a main-effects model is typically assumed (Trocine and Malone 2001, Dean and Lewis 2005).

Other screening methodologies for simulation include one-factor-at-a-time designs (Campolongo et al. 2000); fold-over designs (Myers and Montgomery 2002); methods based on frequency domain analysis (Morrice and Bardhan 1995); edge designs (Elster and Neumaier 1995); iterated fractional factorial designs (Campolongo et al. 2000); and the Trocine screening procedure (Trocine and Malone 2001). These methods will not be discussed in this paper. The interested reader should refer to Trocine and Malone (2000, 2001) or Campolongo et al. (2000) for reviews.

We concentrate on a specific method called *sequential bifurcation* (SB), which is a combination of group screening and a sequential step-down procedure (Bettonvil and Kleijnen 1997). A sequential design is one in which the design points (factor combinations to be studied) are selected as the experiment results become available. Therefore, as the experiment progresses, insight into factor effects is accumulated and used to select the next design point or group of design points.

SB is a series of steps. In each step, the cumulative effect of a group of factors is tested for importance. The first step begins with all factors of interest in a single group and tests that group's effect. If the group's effect is important, indicating that at least one factor in the group may have an important effect, then the group is split into two subgroups. The effects of these two subgroups are then tested in subsequent steps and each subgroup is either classified as unimportant or split into two subgroups for further testing. As the experiment proceeds, the groups become smaller until eventually all factors that have not been classified as unimportant are tested individually. This method was first proposed for deterministic computer simulations by Bettonvil and Kleijnen (1997). Later the method was extended to cover stochastic simulations (Cheng 1997, Kleijnen et al. 2005). Kleijnen et al. (2005) also proposed SB using foldover designs to eliminate the bias of two-factor interactions. The sequential property of the method makes it well suited for simulation experiments. Examples have shown that the method is highly efficient when important factors are sparse and clustered (Cheng 1997, Bettonvil and Kleijnen 1997, Kleijnen et al. 2005), but no one has provided a performance guarantee in the stochastic case.

In this paper, we propose a modified SB procedure, called *controlled sequential bifurcation* (CSB), for stochastic simulations. The contribution of CSB is that it controls the Type I error and power simultaneously. A two-stage testing procedure is introduced to guarantee the power of each step, and at the same time the step-down property of SB implies Type I error control for each factor. The new methodology is an extension of the work of Kleijnen et al. (2005) and Cheng (1997).

This paper is organized as follows: In §2, we define the underlying metamodel that we will use. Section 3 describes the procedure (and a two-stage hypothesis-testing approach) and discusses its performance. In some special situations, a more efficient, fully sequential testing procedure can be implemented, which is discussed in §4. Section 5 presents an empirical evaluation comparing CSB to a number of competitors. In §6, CSB is implemented to solve a realistic problem. Section 7 provides concluding remarks.

2. Response Model

In this section, we introduce the underlying response model that will guide our new CSB procedure.

2.1. Main-Effects Model

Suppose that there are K factors in the simulation experiment. The simulation output of interest is denoted by Y, and Y is represented by the following metamodel:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 z_1 + \tilde{\beta}_2 z_2 + \cdots + \tilde{\beta}_K z_K + \varepsilon, \tag{1}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_K)$ are the K factors, and $\tilde{\beta} = \{\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_K\}$ are the effect coefficients. This is a multiple linear regression model with $K + 1$ regression variables including a dummy variable $z_0 = 1$ and main effects only. The setting of the factors is deterministic and under the control of the experimenter. On the other hand, the error term, ε, is a random variable; in this paper, we assume that it is a $\text{Nor}(0, \sigma^2(\mathbf{z}))$ random variable, where $\sigma^2(\mathbf{z})$ is unknown and may depend on $\mathbf{z}$.

There are two situations in which the main-effects model is appropriate. When there is little prior knowledge about the system and a gross level of screening is desired, then all factors are varied across extremes of their range of operability. In this case, all factors identified as important will be carried to the second stage for a more detailed study and a main-effects model is usually sufficient to identify the candidates for further analysis. However, it should be noted that the main-effects model usually does not hold across the entire range of the factors, so it is possible to miss factors that have large interactions with other factors or have a nonlinear effect on the response.

On the other hand, when the goal of screening is to identify which factors have important local effects (one form of sensitivity analysis), a small disturbance to the nominal level of each factor will be introduced. In this case, the main-effects model is often a good local approximation for modest deviations from a nominal level, typically the center of the design space. CSB is appropriate for both types of screening, but our presentation will focus on the latter application.

2.2. Determination of Factor Levels

In practice, when we consider whether a change in the response is worth pursuing, the cost to achieve the change is often critical. In global screening experiments, management may account for costs by selecting low and high settings of each factor to insure that the ranges of the different factors are comparable. CSB will then code the low settings as zeroes and the high settings as ones (Box and Draper 1987, Chapter 4), and the rest of this section is not relevant. In sensitivity analysis, however, when we compare the effects of two different factors, the comparison may have little meaning if the cost to change the factors is very different. By scaling the effect coefficients with respect to the cost of changing the factors' settings, we can insure that the results have a useful interpretation. In other words, the disturbance to the nominal level for each factor will depend on the cost to change the factor. We describe one way to account for costs here.

Let c_i be the cost per unit change of factor i for $i = 1, 2, \ldots, K$. Further, let $c^* = \max_{i\in\mathscr{D}} c_i$, where $\mathscr{D}$ is the set of indices of all of the factors whose settings can only be changed in discrete units (e.g., number of machines at a workstation or number of cashiers at the checkout). Let Δ_0 be the minimum change in the expected response for which we would be willing to spend c^*, and let Δ_1 be a change in the expected response that we would not want to miss if it could be achieved for only a cost of c^*. If $\mathscr{D} = \varnothing$, then let (c^*, Δ_0) be such that we are willing to spend c^* for a Δ_0 change in the expected response, and define Δ_1 as before.

Let

$$\delta_i = \begin{cases} c^*/c_i, & i \notin \mathscr{D}, \\ \lfloor c^*/c_i \rfloor, & i \in \mathscr{D}, \end{cases}$$

which is the maximum change in factor i that can be achieved without exceeding a cost c^*; and let $w_i = \delta_i c_i/c^* \leqslant 1$, which is the fraction of a full-cost move, c^*/c_i, that can actually be made for factor i. If factor i can be changed continuously ($i \notin \mathscr{D}$), or $i \in \mathscr{D}$ but c^*/c_i is an integer, then $w_i = 1$. If $i \in \mathscr{D}$ and c^*/c_i is not an integer, then $w_i < 1$.

For instance, suppose that there are $K = 3$ factors. The setting of the first factor can be changed continuously, but the other two are discrete. If $c_1 = 300$, $c_2 = 400$, and $c_3 = 1{,}000$, then $c^* = 1{,}000$, $\delta_1 = 10/3$, $\delta_2 = 2$, and $\delta_3 = 1$, giving $w_1 = 1$, $w_2 = 0.8$, and $w_3 = 1$.

Recall that the main-effects model is

$$Y = \tilde{\beta}_0 + \sum_{i=1}^{K} \tilde{\beta}_i z_i + \varepsilon.$$

For screening with a main-effects model, a two-level experimental design is adequate. Let the nominal (low) setting of z_i be z_i^0 and let the high setting be $z_i^0 + \delta_i$ for $i = 1, 2, \ldots, K$. Define the transformed variables $x_i = w_i(z_i - z_i^0)/\delta_i = (c_i/c^*)(z_i - z_i^0)$. Then, Y can be expressed as a linear regression on x_i, $i = 1, 2, \ldots, K$, as

$$Y = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \varepsilon, \tag{2}$$

where the low setting of x_i is 0, the high setting is w_i, and $\beta_i = \delta_i \tilde{\beta}_i / w_i$ for $i = 1, 2, \ldots, K$. Now each β_i, $i > 0$, has a practical interpretation: it represents the change in the expected response when spending c^* to change the setting of factor i, and this change can be compared with Δ_0 and Δ_1 (the thresholds of importance) without ambiguity.

The integration of cost and thresholds of importance into the factor scaling is a general methodology which can be used for any screening strategy. If, on the other hand, the experimenter already knows the thresholds of importance as well as the factor levels, then they do not need to use the cost model. The CSB procedure described in this paper is independent of how the factor levels and thresholds of importance are determined.

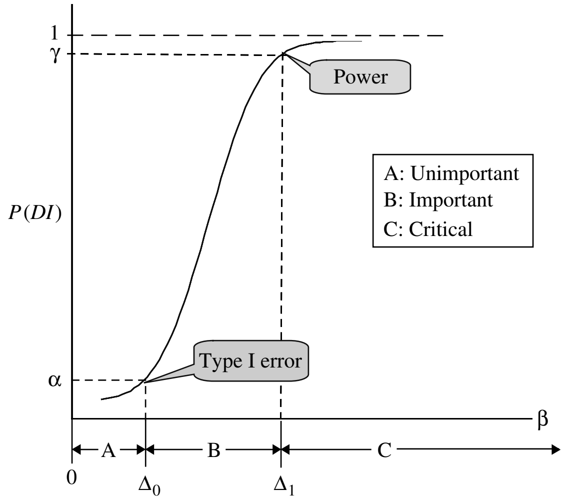
In our proposed method, we will assume that the sign of each factor effect is known so that we can set the levels of the factors to have $\beta_i \geqslant 0$ for all $i > 0$. Further, we assume that for a fixed factor setting, $x_1, x_2, \ldots, x_K$, replications of Model (2) are independent and identically distributed (i.i.d.); dependence of outputs across different factor settings due to CRN is permitted.

2.3. Objective of the Screening Procedure

In screening experiments, the primary objective is to divide the factors into two groups: those that are unimportant, which we take to mean $\beta_i \leqslant \Delta_0$, and those that are important, meaning $\beta_i > \Delta_0$. Because we can never make these determinations with certainty in a stochastic simulation, we instead pursue a screening procedure that controls the probability of incorrectly classifying each factor. More specifically, for those factors with effects $\leqslant \Delta_0$, we require the procedure to control the probability of declaring them important (Type I error) to be less than or equal to α; and for those factors with effects $\geqslant \Delta_1$, we require the procedure to provide power for identifying them as important to be greater than or equal to γ. Here, α and γ are user-specified parameters and Δ_0 and Δ_1 are defined as in §2.2 with $\Delta_1 > \Delta_0$. Those factors whose effects fall between Δ_0 and Δ_1 are considered important and we want the procedure to have reasonable, although not guaranteed, power to identify them. Figure 1 is a generic illustration of the desired performance of our screening procedure. In the figure, the y-axis is $P(DI)$, the probability of declaring a given factor important, and the x-axis is the size of the effect, β.

To illustrate, consider a simulated manufacturing system where the response is the expected throughput of the system. The controllable factors may include the number of machines at each workstation, average processing time of each machine, and the skill levels of the workers. The practical threshold Δ_0 is set as the minimum change in expected throughput that managers consider worth pursuing at a cost c^* of changing the most expensive factor by one unit. For example, c^* might be the cost of purchasing a very expensive machine. In this illustration, screening experiments would be used to identify each factor that influences the expected throughput by more than Δ_0 when spending c^* to

Figure 1. Illustration of desired performance of generic screening procedures.



change that factor. For each factor, the procedure should have probability less than or equal to α of declaring that factor important if it cannot influence the expected throughput by at least Δ_0 at a cost of c^*. The procedure should also have probability greater than or equal to γ of identifying a factor as important if its influence on the expected throughput is greater than or equal to Δ_1 at a cost of c^*. Here, Δ_1 is a critical change in the expected throughput that the managers do not want to ignore if it can be achieved for a cost of only c^*. Factors whose effects are neither unimportant nor critical will be identified with less power than γ.

3. Controlled Sequential Bifurcation

Our CSB procedure inherits the basic structure from the SB procedure proposed by Bettonvil and Kleijnen (1997), and addresses the same problem as the SB-under-uncertainty procedure proposed by Cheng (1997). Specifically, like other SB procedures, CSB is a series of steps in which groups of factors are tested. If a group of factors is considered unimportant, then every factor in the group will be considered unimportant. If the group is considered important, then it is split for further testing. When the algorithm stops, each of the original K factors will be classified as either important or unimportant. The unique feature of CSB is that each step contains a testing procedure to insure the desired power. In addition, CSB preserves the step-down nature of SB so that Type I error can be controlled (a property not noted in previous research on SB; see Hochberg and Tamhane 1987, Theorem 2.6, p. 370). The two-stage procedure is explained in detail in the following sections.

3.1. Notation

The notation that we use to define CSB is given here. It is slightly different from that used in Bettonvil and Kleijnen (1997). (See Appendix A in the online companion at http://or.pubs.informs.org/Pages.collect.html for a complete list of notation used in this paper.)

There are in total K indexed factors. Let x_i represent the setting of factor i. An experiment at level k is defined by the following factor settings, where $w_i = \delta_i c_i / c^* \leqslant 1$:

$$x_i(k) = \begin{cases} w_i, & i = 1, 2, \ldots, k, \\ 0, & i = k+1, k+2, \ldots, K. \end{cases}$$

Thus, "level k" indicates an experiment at which factors $1, 2, \ldots, k$ are set at their high settings, and factors $k+1$, $k+2, \ldots, K$ are set at their low settings. Note that in this paper, "setting" is used to describe a single factor's value; and "level" is used to describe all factors' settings in an experiment.

Let $Y_j(k)$ denote the jth simulation replication of an experiment at level k. Therefore, under our main-effects model, $Y_j(k) = \beta_0 + \sum_{i=1}^{k} w_i \beta_i + \varepsilon_j(k)$. We require $\varepsilon_j(k)$ to be independent of $\varepsilon_{j'}(k')$ when $j \neq j'$ (i.i.d. assumption in §2.2). In the case of CRN, there will be positive correlation between $\varepsilon_j(k)$ and $\varepsilon_j(k')$, $k \neq k'$, which enables more precise estimators of the effects.

When a level is selected for observation, N_0 replications will be initially taken, but more generally, n_k denotes the number of replications that have been taken at level k. For $k_2 > k_1$, let $D_j(k_1, k_2) = Y_j(k_2) - Y_j(k_1)$, $j = 1, 2, \ldots,$ $\min\{n_{k_1}, n_{k_2}\}$, and let the average of the differences of the paired observations be

$$\bar{D}(k_1, k_2) = \sum_{j=1}^{\min\{n_{k_1}, n_{k_2}\}} (Y_j(k_2) - Y_j(k_1)) / \min\{n_{k_1}, n_{k_2}\},$$

whose expected value is $\sum_{i=k_1+1}^{k_2} w_i \beta_i$. The quantity $\bar{D}(k_1, k_2)$ will be our test statistic for determining whether or not the group $\{\beta_{k_1+1}, \beta_{k_1+2}, \ldots, \beta_{k_2}\}$ is important. However, because different factors may have different weights, we use $w(k_1, k_2) = \min\{w_{k_1+1}, w_{k_1+2}, \ldots, w_{k_2}\}$, the smallest weight associated with factors $\beta_{k_1+1}, \beta_{k_1+2}, \ldots, \beta_{k_2}$, to scale the test statistic. Our test also requires

$$S^2(k_1, k_2) = \frac{1}{N_0 - 1} \sum_{j=1}^{N_0} (D_j(k_1, k_2) - \bar{D}(k_1, k_2))^2,$$

the first-stage (initial N_0 observations) sample variance of the paired differences. The critical regions for CSB's hypothesis tests are defined by the following quantities:

- $t_{\theta, \nu}$: the θ quantile of the t-distribution with ν degrees of freedom.
- $U_A(k_1, k_2) = \Delta_0 + t_{\sqrt{1-\alpha}, N_0-1} S(k_1, k_2) / (w(k_1, k_2)\sqrt{n_k})$, where $n_k = \min\{n_{k_1}, n_{k_2}\}$. The subscript $A = \text{I, II}$ denotes the first or second stage of the testing procedure, respectively.
- $L(k_1, k_2) = \Delta_0 - t_{(1+\gamma)/2, N_0-1} S(k_1, k_2) / (w(k_1, k_2)\sqrt{n_k})$, where $n_k = \min\{n_{k_1}, n_{k_2}\}$.
- h: A constant such that $\mathrm{P}(T_{N_0-1} \leqslant t_{\sqrt{1-\alpha}, N_0-1} - h) = (1+\gamma)/2$, where T_{N_0-1} is a t-distributed random variable with $N_0 - 1$ degrees of freedom.
- $N(k_1, k_2) = \lceil h^2 S^2(k_1, k_2) / (w^2(k_1, k_2)(\Delta_1 - \Delta_0)^2) \rceil$, the total sample size at the end of Stage II.

Figure 2. Structure of CSB.

```
Initialization: Create an empty LIFO queue for groups.
  Add the group $\{1, 2, \ldots, K\}$ to the LIFO queue.
While queue is not empty, do
  Remove: Remove a group from the queue.
  Test:
    Unimportant: If the group is unimportant, then classify all
      factors in the group as unimportant.
    Important (size $= 1$): If the group is important and of size 1,
      then classify the factor as important.
    Important (size $> 1$): If the group is important and size is
      greater than 1, then split it into two subgroups such that all
      factors in the first subgroup have smaller index than those
      in the second subgroup. Add each subgroup
      to the LIFO queue.
  End Test
End While
```

In the next section, we show how the test is performed.

3.2. CSB Procedure

An overview description of CSB is shown in Figure 2. The figure illustrates how groups are created, manipulated, tested and classified, but does not specify how data are generated or what tests are performed. Detailed descriptions of data collection and hypothesis testing follow. This section is closed by an example.

Independent and identically distributed replications are obtained whenever new groups are formed according to the following rule: When forming a new group containing factors $\{k_1+1, k_1+2, \ldots, k_2\}$ with $k_1 < k_2$, check the number of replications at levels k_1 and k_2:

If $n_{k_1} = 0$, then collect N_0 replications at level k_1 and set $n_{k_1} = N_0$.

If $n_{k_2} = 0$, then collect N_0 replications at level k_2 and set $n_{k_2} = N_0$.

If $n_{k_1} < n_{k_2}$, then make $n_{k_2} - n_{k_1}$ additional replications at level k_1 and set $n_{k_1} = n_{k_2}$.

If $n_{k_2} < n_{k_1}$, then make $n_{k_1} - n_{k_2}$ additional replications at level k_2 and set $n_{k_2} = n_{k_1}$.

Suppose that the group removed from the queue contains factors $\{k_1+1, k_1+2, \ldots, k_2\}$ with $k_1 < k_2$. The **Test** step in Figure 2 tests the following hypothesis to determine if this group might contain important factors:

$$\mathrm{H}_0\colon \sum_{i=k_1+1}^{k_2} \beta_i \leqslant \Delta_0 \quad \text{vs.} \quad \mathrm{H}_1\colon \sum_{i=k_1+1}^{k_2} \beta_i > \Delta_0.$$

The procedure given below for testing this hypothesis guarantees that the probability of Type I error is less or equal to α when $\sum_{i=k_1+1}^{k_2} \beta_i \leqslant \Delta_0$, and the power is greater or equal to γ if $\sum_{i=k_1+1}^{k_2} \beta_i \geqslant \Delta_1$.

Two-Stage Test

Stage I

1. If $\bar{D}(k_1, k_2)/w(k_1, k_2) \leqslant U_{\mathrm{I}}(k_1, k_2)$ and $\min\{n_{k_1}, n_{k_2}\} \geqslant N(k_1, k_2)$, then classify the group as unimportant.
2. Else if $\bar{D}(k_1, k_2)/w(k_1, k_2) \leqslant L(k_1, k_2)$, then classify the group as unimportant.
3. Else if $\bar{D}(k_1, k_2)/w(k_1, k_2) > U_{\mathrm{I}}(k_1, k_2)$, then classify the group as important.
4. Else go to Stage II.

Stage II

5. Make $(N(k_1, k_2) - n_{k_1})^+$ replications at levels k_1 and k_2 (recall that $n_{k_1} = n_{k_2}$). Then set $n_{k_1} = n_{k_2} = \max\{N(k_1, k_2), n_{k_1}\}$. The sample variance $S^2(k_1, k_2)$ and the degrees of freedom do not change, but $\bar{D}(k_1, k_2)$ is updated.

(a) If $\bar{D}(k_1, k_2)/w(k_1, k_2) \leqslant U_{\mathrm{II}}(k_1, k_2)$, then classify the group as unimportant.

(b) If $\bar{D}(k_1, k_2)/w(k_1, k_2) > U_{\mathrm{II}}(k_1, k_2)$, then classify the group as important.

Note that because $w_i \leqslant 1\ \forall\, i$,

$$\mathrm{E}[\bar{D}(k_1, k_2)] = \sum_{i=k_1+1}^{k_2} w_i \beta_i \leqslant \sum_{i=k_1+1}^{k_2} \beta_i.$$

Therefore testing $\bar{D}(k_1, k_2)$ against Δ_0 would sacrifice power and be conservative for Type I error. Thus, we use $\bar{D}(k_1, k_2)/w(k_1, k_2)$ because $\mathrm{E}[\bar{D}(k_1, k_2)/w(k_1, k_2)] \geqslant \sum_{i=k_1+1}^{k_2} \beta_i$. Type I error is controlled by testing singleton groups in the final steps. Because $\mathrm{E}[\bar{D}(k_1, k_2)/w(k_1, k_2)] = \beta_{k_2}$ when $k_1 + 1 = k_2$, experimentwise Type I error control will not be compromised even though the Type I error for nonsingleton group testing is no longer conservative (see Appendix B in the online companion at http://or.pubs.informs.org/Pages.collect.html).

As an illustration, consider the case of $K = 10$ factors and the first pass through the algorithm. Initially, we make N_0 replications at level 0 (all factors at their low settings) and N_0 replications at level 10 (all factors at their high settings). The group removed from the queue contains all factors and $w(0, 10) = \min\{w_1, w_2, \ldots, w_{10}\}$.

Next, we evaluate $\bar{D}(0, 10)$, $U_{\mathrm{I}}(k_1, k_2)$, and $L(k_1, k_2)$. If $\bar{D}(0, 10)/w(0, 10) \leqslant L(k_1, k_2)$, then we conclude that none of the factors is important because the sum of all effects is not important, and the algorithm stops. If $\bar{D}(0, 10)/w(0, 10) > U_{\mathrm{I}}(k_1, k_2)$, then the factors are separated into two groups, $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ and $\{\beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}\}$, and N_0 replications are made at level 5 (which means that x_i, $i = 1, 2, \ldots, 5$, are at their high settings and x_i, $i = 6, 7, \ldots, 10$, are at their low settings). Both groups are added to the queue.

If, on the other hand, $\bar{D}(0, 10)/w(0, 10)$ is between $L(k_1, k_2)$ and $U_{\mathrm{I}}(k_1, k_2)$, then we calculate $N(0, 10)$. If $N(0, 10) \leqslant N_0$, then we conclude that no factors are important and the algorithm stops. If $N(0, 10) > N_0$, then we collect $N(0, 10) - N_0$ replications at both level 0 and level 10, reevaluate $\bar{D}(0, 10)$, and calculate $U_{\mathrm{II}}(k_1, k_2)$. If $\bar{D}(0, 10)/w(0, 10) > U_{\mathrm{II}}(k_1, k_2)$, then the factors are separated into two groups as described above and N_0 replications are made at level 5. Both groups are added to the queue. Otherwise, all factors will be considered to be unimportant and the algorithm stops.

3.3. Implementation Issues

The following are key issues in our implementation of CSB. Other choices could be made, and these are the subject of future research.

Group Splitting. Our current version of CSB splits an important group in the middle. When the number of factors in the group is odd, the group containing factors with smaller indices will get one more factor. So for a group containing factors $\{k_1+1, k_1+2, \ldots, k_2\}$ with $k_1 < k_2$, we split at the point $k = \lceil (k_1+k_2)/2 \rceil$ and the two new groups contain factors $\{k_1+1, k_1+2, \ldots, k\}$ and $\{k+1, k+2, \ldots, k_2\}$, respectively. There are other policies available (see Kleijnen et al. 2005).

Number of Replications at Each Level. In our current version of CSB, we always make enough replications to insure that $n_{k_1} = n_{k_2}$ before performing the hypothesis test. This is to make sure that the response from high and low levels of the testing group are paired, which is mandatory when CRN is implemented.

If we are willing to store each observation instead of merely their summary statistics, a modified strategy is to conduct the test based on the first $\min\{n_{k_1}, n_{k_2}\}$ paired observations from low and high levels. If it turns out that more observations are needed, take $(N(k_1,k_2) - n_{k_1})^+$ and $(N(k_1,k_2) - n_{k_2})^+$ more observations from low and high levels, respectively. This usually decreases the total number of replications required for screening, but also increases the memory requirement.

Indexing of the Factors. It is preferable to index the factors monotonically by their effect size so that the unimportant factors are likely to be eliminated together at early stages (Bettonvil and Kleijnen 1997). However, most of the time the relative size of the effects is not known. Another option is to have the factors indexed monotonically by w_i so that the small w_i are grouped together. Consider a group containing factors $\{k_1+1, k_1+2, \ldots, k_2\}$, $0 \leqslant k_1 < k_2 \leqslant K$. It is desirable to have the w_i in the group as close in value as possible so that $\mathrm{E}[\bar{D}(k_1,k_2)/w(k_1,k_2)]$ will be close to $\sum_{i=k_1+1}^{k_2} \beta_i$. Both strategies improve the efficiency of the procedure, but they may be in conflict. If we suspect that there are large differences between effect sizes and we know the order, then the factors should be indexed by their effect size. When we have no such knowledge (the usual case), then the factors should be indexed by w_i.

Early Stopping. Because both the number of steps required for screening and the number of replications required at each step are automatically adjusted in CSB, the experimenter has little control over the computational effort that will be used for screening. However, early stopping is possible when time or resources are limited. If stopped in the middle of the screening process, CSB will give one or more subgroups, possibly containing important factors, and a group of factors that have already been classified as unimportant. The unclassified factors can be carried to the next stage for further testing. Specifically, if the number of factors we are willing to carry to the next stage has been pre-specified as K', then as soon as $K - K'$ (or more) factors have been eliminated, the CSB procedure can stop. Early termination can save substantially on the number of runs required.

3.4. Performance of CSB

The performance guarantees for the CSB procedure are stated in following theorems that are proved in Appendix B in the online companion.

Theorem 1. *If Model* (2) *holds with normally distributed error and all* $\beta_i \geqslant 0$, $i > 0$, *then CSB guarantees that*

$$\mathrm{P}_{\beta_i \leqslant \Delta_0}\{\textit{declare factor } i \textit{ important}\} \leqslant \alpha$$

for each factor i individually.

Theorem 2. *Let the group containing the factors denoted* $\{k_l+1, \ldots, k_m\}$ *be represented by* $\{k_l \to k_m\}$ *for* $0 \leqslant k_l < k_m \leqslant K$. *If Model* (2) *holds with normally distributed error and all* $\beta_i \geqslant 0$, $i > 0$, *then the two-stage test guarantees that*

$$\mathrm{P}_{\sum_{i=k_l+1}^{k_m} \beta_i \geqslant \Delta_1}\{\textit{declare } \{k_l \to k_m\} \textit{ important}\} \geqslant \gamma$$

for each group $\{k_l \to k_m\}$ *tested.*

In summary, the CSB procedure controls the Type I error for each factor individually and guarantees the power for each step. The procedure does not require an equal variance assumption, and is valid with or without CRN. At the end of the procedure the factors are separated into two groups, those that are classified important and those that are classified not important. For each unimportant factor, the probability that it will be classified as important is less than or equal to α. The power to detect effects greater than or equal to size Δ_1 (the critical effects) is controlled at each testing step, but not experimentwise. However, because in all but the last step, $\sum_{i=k_l+1}^{k_m} \beta_i$ is likely to be greater than Δ_1 if any one factor has an effect of size Δ_1, the power should not be seriously compromised. An empirical evaluation will be discussed in §5.

More generally, if CSB employs any testing procedure that guarantees:

1. $\mathrm{P}_{\sum_{i=k_l+1}^{k_m} \beta_i \leqslant \Delta_0}$\{declare $\{k_l \to k_m\}$ important\} $\leqslant \alpha$; and
2. $\mathrm{P}_{\sum_{i=k_l+1}^{k_m} \beta_i \geqslant \Delta_1}$\{declare $\{k_l \to k_m\}$ important\} $\geqslant \gamma$,

then the conditions of Theorems 1 and 2 will still be satisfied. On the other hand, if a testing procedure only fulfills requirement 1, then Theorem 1 will hold, but we lose the power control. For example, if we eliminate the second stage of the two-stage testing procedure described above, and make only a one-sided hypotheses test in the first stage, then the procedure is very similar to the approach of Kleijnen et al. (2005). Theorem 1 will still hold, but Theorem 2 will not.

The Type I error control, described in Theorem 1, is for each factor individually. We will briefly discuss the experimentwise Type I error control of CSB by evaluating the expected number of factors that are falsely classified as important, denoted $\mathrm{E}[F_K]$, for two extreme cases.

To simplify the analysis, we assume that there are $K = 2^L$ factors, where L is an integer. Therefore, CSB needs L tests to get down to a group of size 1 (a singleton group), and a singleton group must be declared important for a factor to be declared important. Also, we assume that all tests are independent. Then, we have the following two theorems:

Theorem 3. *If Model* (2) *holds with normally distributed error, all* $\beta_i \geqslant 0$, $i > 0$, *and* $\sum_{i=1}^{K} \beta_i \leqslant \Delta_0$, *then for* $\alpha \leqslant 1/2$, *CSB guarantees that*

$$\mathrm{E}[F_K] \leqslant \alpha.$$

In this case, no factor or group of factors is significant. The upper bound for $\mathrm{E}[F_K]$ is scale-free and the bound decreases with decreasing α.

Theorem 4. *If Model* (2) *holds with normally distributed error, all* $\beta_i \geqslant 0$, $i > 0$, *and* $\beta_i \leqslant \Delta_0$, $i = 1, 2, \ldots, K$, *but* $\beta_i + \beta_j \geqslant \Delta_1$ *for all* $i \neq j$, *then CSB guarantees that*

$$\mathrm{E}[F_K] \leqslant K\alpha.$$

Theorem 4 examines the worst case for controlling the Type I error because all factors should be carried to the last step and tested separately, but none of the factors is important. The upper bound for $\mathrm{E}[F_K]$ is linear in the number of factors. Realistic problems should be between these two extreme cases, but closer to Theorem 3. Therefore, CSB provides strong control of the "false positive" rate, regardless of the number of factors.

4. CSB with a Fully Sequential Testing Procedure

In the two-stage testing procedure introduced in §3, the determination of the second-stage sample size is based on a worst-case scenario. Specifically, the test assumes that all important effects are on the boundary of critical importance, Δ_1. Even if the effect size is much larger than Δ_1, the test requires just as much data to guarantee the power. In the special case where $\alpha = 1 - \gamma$ (Type I error is equal to one minus the power), a fully sequential test can be implemented in CSB that gives the same error control as the two-stage testing procedure. The test adds one replication at a time to both the upper and lower levels of the group being tested until a decision is made. In most cases, the sequential test is more efficient than the two-stage testing procedure. The test is adapted from Kim (2005), and we refer the reader to that paper for the proof of its validity. When we use this test, the structure of CSB remains the same as in Figure 2; only the **Test** step is changed.

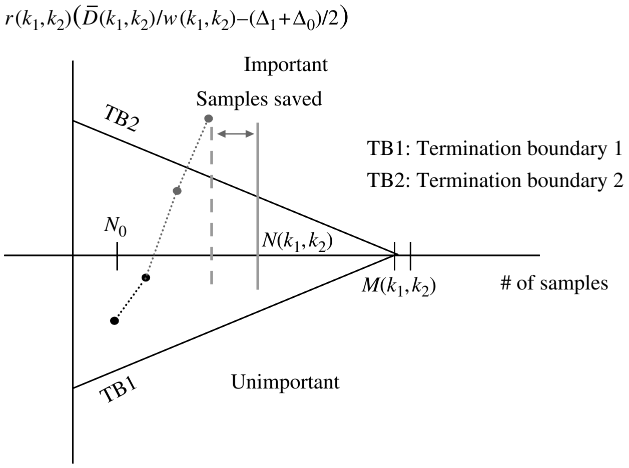
Kim's sequential test has three critical constants which in our context are set as follows:

$\eta = (\exp\varphi - 1)/2$ with $\varphi = -2\ln(2\alpha)/(N_0 - 1)$,

$a(k_1, k_2) = 2\eta(N_0 - 1)S^2(k_1, k_2)/(w^2(k_1, k_2)(\Delta_1 - \Delta_0))$,

$\lambda = (\Delta_1 - \Delta_0)/4$.

Figure 3. Fully sequential test.


$r(k_1,k_2)(\bar{D}(k_1,k_2)/w(k_1,k_2)-(\Delta_1+\Delta_0)/2)$
Important
Samples saved
TB2
TB1: Termination boundary 1
TB2: Termination boundary 2
N_0
$N(k_1,k_2)$
$M(k_1,k_2)$
# of samples
TB1
Unimportant

Let $r(k_1, k_2)$ denote the current number of replications at levels k_1 and k_2. The fully sequential test takes paired observations from each level, one pair at a time, and checks whether $\bar{D}(k_1, k_2)$ crosses one of two termination boundaries, which are functions of η, $a(k_1, k_2)$, and λ. The maximum number of paired observations that will be taken is one more than $M(k_1, k_2) = \lfloor a(k_1, k_2)/\lambda \rfloor$. As illustrated in Figure 3, if the group effect is significantly larger (or smaller) than the threshold of importance, the conclusion can be made with a smaller number of observations than the two-stage testing procedure which requires $N(k_1, k_2)$. In the figure, the dots represent the value of the test statistic as a function of the number of paired observations. Note that $N(k_1, k_2) \leqslant M(k_1, k_2)$, meaning the maximum number of observations the fully sequential test could take is greater than the number of observations the two-stage test will take. If the effect is on the boundary of importance, it is possible that the fully sequential test will continue until $M(k_1, k_2) + 1$ observations have been collected. In this case, the test is not as efficient. Fortunately, the former case usually happens more often than the latter, which makes the fully sequential testing procedure more efficient than the two-stage test. After the initial N_0 observations ($r(k_1, k_2) = N_0$), the test works as follows:

Fully Sequential Test

1. If $r(k_1, k_2) > M(k_1, k_2)$, then
   (a) If $r(k_1, k_2)(\bar{D}(k_1, k_2)/w(k_1, k_2) - (\Delta_0 + \Delta_1)/2) \leqslant 0$, then stop and classify the group as unimportant.
   (b) Else stop and classify the group as important.
2. Else (i.e., $r(k_1, k_2) \leqslant M(k_1, k_2)$)
   (a) If $r(k_1, k_2)(\bar{D}(k_1, k_2)/w(k_1, k_2) - (\Delta_0 + \Delta_1)/2) \leqslant -a(k_1, k_2) + \lambda r(k_1, k_2)$ (termination boundary 1), then classify the group as unimportant.
   (b) Else if $r(k_1, k_2)(\bar{D}(k_1, k_2)/w(k_1, k_2) - (\Delta_0 + \Delta_1)/2) \geqslant a(k_1, k_2) - \lambda r(k_1, k_2)$ (termination boundary 2), then classify the group as important.
   (c) Else take one more replication at both levels k_1 and k_2, set $r(k_1, k_2) = r(k_1, k_2) + 1$, and go to Step 1.

This test is a special case of a more general fully sequential ranking-and-selection procedure due to Kim (2005).

Kim's procedure handles the problem of comparing m simulated systems with a standard system. If the designated standard is system 0 and $i = 1, 2, \ldots, m$ are the alternative systems, then the goal is to identify the system with the largest expected performance provided it is significantly better than the standard. Suppose that system i has expected performance μ_i; without loss of generality, we can assume that $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_m$. Kim's procedure guarantees the probability of correct selection to be $\geqslant 1-\alpha$ given a practically significant difference $\delta > 0$ worth detecting:

$$\mathrm{P}\{\text{select system } 0\} \geqslant 1-\alpha \text{ whenever } \mu_0 \geqslant \mu_m,$$

$$\mathrm{P}\{\text{select system } m\} \geqslant 1-\alpha \text{ whenever } \mu_m \geqslant \max\{\mu_{m-1}, \mu_0\} + \delta.$$

We have only two "systems" at each given step. To adapt Kim's procedure to our setting, we identify system 0 with the threshold of importance Δ_0, identify system $m = 1$ with the group effect, set $\delta = \Delta_1 - \Delta_0$, and let $\alpha = 1-\gamma$. If system 0 is "selected," then the group is classified as unimportant; if system 1 is "selected," then the group is classified as important. The performance guarantee of Kim's procedure implies that Lemma 3 of Appendix B in the online companion and Theorems 1–4 still hold. Therefore, CSB with the fully sequential test has the same error control as CSB with the two-stage test discussed in §3.

5. Empirical Evaluation

In this section, we discuss the numerical results that compare CSB with the two-stage testing procedure proposed in §3 to Cheng's method (Cheng 1997), an enhancement of the SB procedure for stochastic responses that assumes equal variances. We also compare the efficiency of the two-stage and fully sequential testing procedures, and demonstrate the advantages of CSB relative to traditional fractional factorial designs when the number of factors is large and the number of important factors is small. Finally, we compare CSB to the SB procedure of Kleijnen et al. (2005).

5.1. Comparison of CSB and Cheng's Method

The idea behind Cheng's (1997) method is to determine whether a group of two or more factors is unimportant by constructing a one-sided confidence interval for the group's effect. For a group containing a single factor, replications are added one at a time until a two-sided confidence interval for the factor effect shows that the effect is important or unimportant. When a single factor is tested, the method employs an indifference parameter a. In our notation, all the factors with effects smaller than $\Delta_0 + a$ can be classified as unimportant. Cheng's method does not guarantee the control of Type I error for each factor or control of the power at any step, and has no concept like Δ_1 for a critically important factor. In this section, the CSB method refers to CSB with the two-stage testing procedure.

5.1.1. Summary of Results. Rather than employ system simulation models in this test, we chose to generate data from a main-effects model in which we control the size of the effects and the variances at different design points; a realistic example is given in §6. Normal errors are assumed with mean 0 and standard deviation $\sigma = m * (1 + \mathcal{I} * \text{size of the group effect})$, where $\mathcal{I}$ is 0 if we are running an equal-variance case, and 1 for an unequal-variance case. Thus, in unequal variance cases, the standard deviation is proportional to the size of the effect of the group being screened. The parameter m determines the magnitude of the variance. CRN were not employed because Cheng's procedure is not valid under CRN because it assumes independence and equal variance for each observation.

For each case considered, the CSB procedure using the two-stage test of §3.2 is applied 1,000 times and the percentage of time factor i declared important is recorded; this is an unbiased estimator of P{factor i is declared important}.

To compare CSB to Cheng's (1997) method, we set the indifference parameter, a, such that the number of replications required by Cheng's method is approximately the same as the number used by CSB for that case. Therefore, we can compare the estimated probability of Type I error and power of the two methods with equal simulation effort.

The performance of Cheng's method depends on the case considered. When the variances are large and unequal, Cheng's method loses control of both the Type I error and the power. The CSB method, on the other hand, controls the Type I error and power across all cases (although the number of replications required to achieve this does differ substantially by case).

In the following subsections, we provide some illustrative numerical results that emphasize the key conclusions.

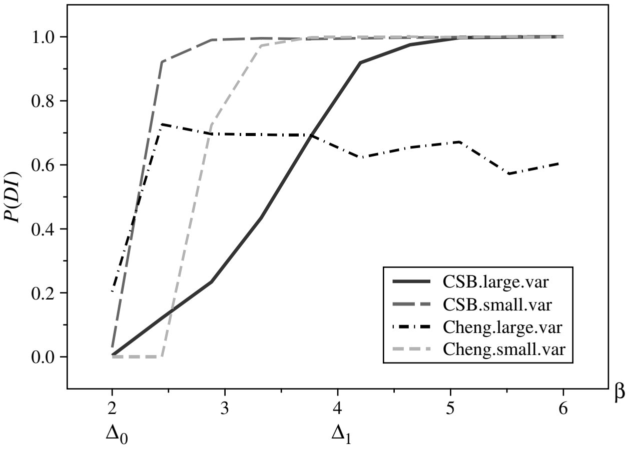
5.1.2. Unequal-Variance Cases. We set the parameters as in Table 1. We considered two different settings for the factor effects:

1. In Case 1, we set $(\beta_1, \beta_2, \ldots, \beta_{10}) = (2, 2.44, 2.88, 3.32, 3.76, 4.2, 4.64, 5.08, 5.52, 6)$, spanning the range from Δ_0 to $\Delta_0 + \Delta_1$. For CSB, the observed frequency that β_1 is declared important should be smaller than 0.05, but for $\beta_6, \beta_7, \ldots, \beta_{10}$ it should be near 0.95. Letting $P(DI)$ mean "probability of being declared important," Figure 4 plots $P(DI)$ against effect size for Cheng's method and

Table 1. Parameters for main effects experiments.

Parameter	Value
K	10
Δ_0	2
Δ_1	4
α	0.05
γ	0.95
m	0.1, 1

Figure 4. Case 1 with unequal variances.



CSB with large ($m = 1$) and small ($m = 0.1$) variances. We can see that when variance is small, the two methods have similar performance although CSB attains greater power earlier. When the variance is large, however, Cheng's method loses control of both Type I error and power.

2. In Case 2, we set $(\beta_1, \beta_2, \ldots, \beta_{10}) = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$, so that all effects are Δ_0. This set is designed to study the control of Type I error for the two methods. The other parameters are the same as in the previous case.

Figure 5 shows the box plots of the Type I error control of both methods. The frequency of Type I error is large for Cheng's method when the variance is large. Even for the small-variance case, the largest frequency of Type I error is still more than 0.2 for Cheng's method. By design, CSB controls the probability of Type I error to be $\leqslant \alpha$ in all cases.

5.1.3. Equal-Variance Cases. The parameter settings are the same as the unequal variance cases except that $\sigma = m$, which is the same across all responses. We considered two different settings for the factor effects:

1. In Case 1, we set $(\beta_1, \beta_2, \ldots, \beta_{10}) = (2, 2.44, 2.88, 3.32, 3.76, 4.2, 4.64, 5.08, 5.52, 6)$. The results are summarized in Figure 6. This time the two methods perform similarly, although CSB has somewhat larger power.

Figure 5. Case 2 with unequal variances.

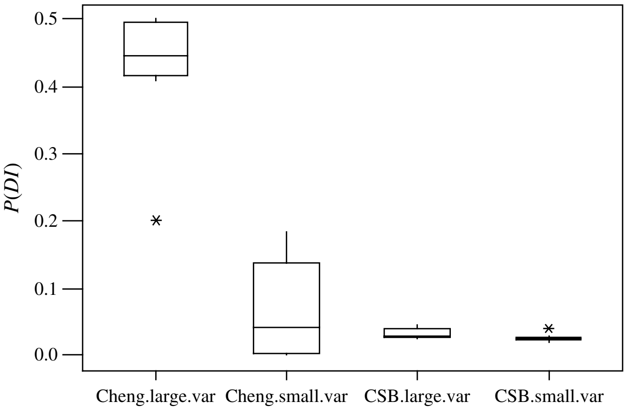
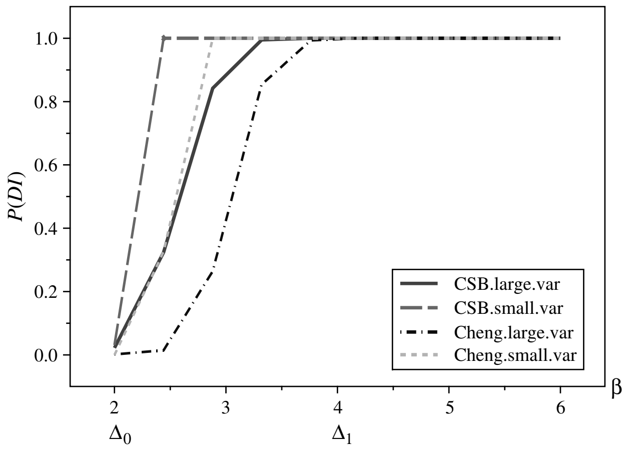


Figure 6. Case 1 with equal variances.



2. In Case 2, we set $(\beta_1, \beta_2, \ldots, \beta_{10}) = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$. As shown in Figure 7, CSB has a better control of Type I error for both large and small variances.

To summarize, CSB has performance superior to Cheng's method in large and unequal variance cases. CSB has guaranteed performance with different parameter and factor configurations, which makes it attractive for problems with limited prior knowledge. Cheng's method, on the other hand, assumes variance homogeneity to gain advantages of increased degrees of freedom and it can be effective when this assumption is satisfied.

5.2. Comparison of Two-Stage and Fully Sequential Tests

CSB with the fully sequential test of §4 gives very similar results for Type I error and power as in §5.1. The difference lies in the number of replications required for screening. To compare the efficiency of the two tests, the experiments presented in §5.1 are repeated with CSB using the fully sequential test, and the average number of replications

Figure 7. Case 2 with equal variances.

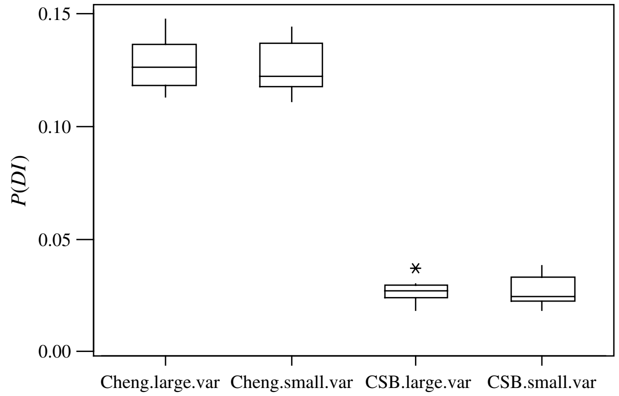


Table 2. Efficiency comparison of CSB with two-stage and fully sequential testing procedure.

Case	Sequential	Two-stage
Unequal variance case 1, $m=1$	13,579	30,397
Unequal variance case 1, $m=0.1$	306	302
Unequal variance case 2, $m=1$	8,947	14,920
Unequal variance case 2, $m=0.1$	285	290
Equal variance case 1, $m=1$	275	275
Equal variance case 1, $m=0.1$	275	275
Equal variance case 2, $m=1$	275	275
Equal variance case 2, $m=0.1$	275	275

required for each case is compared with CSB using the two-stage test. The results are summarized in Table 2.

We can see that CSB using the fully sequential test is more efficient in the unequal variance cases, especially when the variance is large. For the particular equal variance cases in these examples, the number of replications for each test is never more than the initial N_0 replications so the two tests have the same performance.

Although CSB using the fully sequential test typically takes fewer replications than CSB using the two-stage test, it does not necessarily mean that the fully sequential testing procedure is always preferred. The cost of continually switching between the settings of the factors can be high, especially if done manually. The two-stage testing procedure is also simpler to implement and does not require $\alpha = 1-\gamma$.

5.3. Comparison of CSB and Fractional Factorial Design for Large-Scale Problems

As discussed previously, CSB takes advantage of the highly sequential nature of simulation experiments and is more efficient than traditional methodologies when the number of factors is large and only a small fraction of them is important. In this section, we study screening problems with 200 factors and 500 factors and compare CSB to a standard unreplicated fractional factorial design (i.e., an orthogonal array). For each case, only 2% of the factors are important. The important factors have effects equal to 5 and the unimportant factors have effects equal to 0. Normal errors are assumed with mean 0 and standard deviation 1 (equal variance across different levels). The threshold of importance, Δ_0, is set to 2; and the critical threshold, Δ_1, is set to 4. The initial number of runs at each level, N_0, is equal to 5 for the 200 factor case and 8 for the 500 factor case. The Type I error is set to be $\alpha = 0.05$ and the power requirement is $\gamma = 0.95$.

For each case, there are two scenarios. The first scenario has all important factors clustered together with the smallest indices so that the number of important groups is as small as possible at each step. The second scenario has the important factors evenly spread so there are the maximum number of important groups remaining at each step. CSB is more efficient with the first scenario than with the second scenario.

Table 3. Comparison of CSB and fractional factorial design.

	Number of runs required	
Scenarios	CSB	FFD
200 factors, $\{1,2,3,4\}$ Important	79	256
200 factors, $\{1,51,101,151\}$ Important	282	256
500 factors, $\{1,2,3,4,5,6,7,8,9,10\}$ Important	148	512
500 factors, $\{1,51,101,151,201,251,301,351,401,451\}$ Important	573	512

For each case and scenario considered, CSB with the fully sequential testing procedure is applied 1,000 times and the average number of replications required for screening is recorded. The number of replications required for the fractional factorial design is the number of design points required to estimate 200 or 500 main effects with a Resolution III design, which is not influenced by the scenario. The comparison is shown in Table 3, where FFD represents fractional factorial design.

We can see that for both the 200 and 500 factors cases, CSB only takes approximately 1/4 to 1/3 of the replications required by fractional factorial designs in the clustered scenario. In the other scenario, the fractional factorial design requires fewer replications, but the difference is small. Realistic problems are usually between the two scenarios. Thus, CSB is typically more efficient. In addition, if the conditions of Theorem 3 or Theorem 4 are satisfied, the expected number of factors that will be falsely classified as important ($\mathrm{E}[F_K]$) for the fractional factorial design always equals αK, which is greater than or equal to that of CSB, especially for the conditions of Theorem 3. Furthermore, the fractional factorial design does not have power control. Therefore, from the error-control point of view, CSB is superior. Moreover, the typical test implemented in a fractional factorial design assumes equal variance across design points, which CSB does not require. However, the fractional factorial design does not need $\beta_i \geqslant 0$ $(i>0)$.

5.4. Comparison of CSB to Sequential Bifurcation with One-Stage Hypothesis Testing Procedure

Kleijnen et al. (2005) have proposed using SB with a one-stage hypothesis testing procedure (SB-One). At a bifurcation step, suppose that the testing group contains factors $\{k_1+1, k_1+2, \ldots, k_2\}$ with $k_1 < k_2$, and define $L'(k_1,k_2) = \Delta_0 + t_{1-\alpha, N_0-1} S(k_1,k_2)/(w(k_1,k_2)\sqrt{N_0})$. If $D(k_1,k_2)/w(k_1,k_2) \leqslant L'(k_1,k_2)$, the group will be classified as unimportant; otherwise the group is split into two subgroups for further testing, unless the group is singleton, in which case the factor will be classified as important. This procedure is equivalent to CSB without the subsequent stage(s) of sampling to control the power. An upper bound on the number of runs required for SB-One is $(K+1)N_0$. The Type I error control for each factor still holds because each bifurcation step controls the Type I error. This property is not mentioned in Kleijnen et al. (2005) but clearly follows from our Theorem 1. On the other hand, power is not controlled.

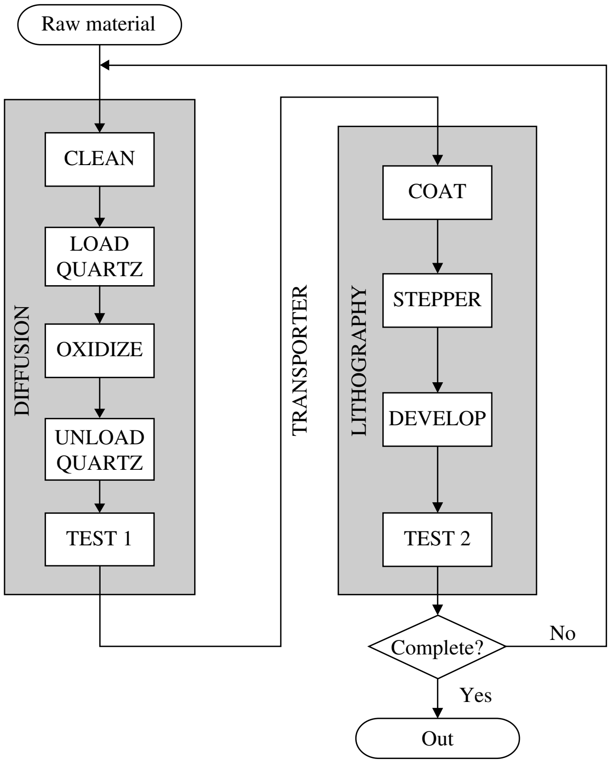
We performed the same numerical study described in §5.1 using SB-One. When variances are relatively small (Cases 1 and 2 with $m = 0.1$), the performance of SB-One is similar to CSB. However, if variances are relatively large ($m = 1$), then for Case 1 where factor effects increase gradually, SB-One loses control of power for large effects. For example, $\mathrm{P}(DI) = 0.06$ when $\beta = 6$, even though the size of the effect is very large relative to Δ_0.

In summary, incorporating a hypothesis test into SB only guarantees control of Type I error; to control power and mitigate the risk of missing important factors, an adaptive test, such as our two-stage and fully sequential tests, is required. The same can be said for other methods (factorial design, for example) without explicit control of power.

6. Case Study

The case study discussed in this section is a simplified and modified version of a real-world problem. Consider a semiconductor manufacturing system. In our simplified version, the production process consists of two basic steps, diffusion and lithography, each of which contains substeps as indicated in Figure 8.

Figure 8. Production process of the semiconductor manufacturing case study.



Raw material will be released in cassettes at the rate of one cassette per hour, seven days per week, eight hours per day. Product is moved and processed in single cassette loads. The raw material will begin at the diffusion process, and after diffusion it proceeds to the lithography process. The diffusion and lithography then alternate until the product completes processing. Movement within each process is handled by robots; the handling time is negligible. The movement of material from the end of diffusion to the start of lithography (or vice versa) will be handled by an automatic guided vehicle (AGV) or a conveyor. The release of the raw material, the processing of material at each station, and the transportation between diffusion and lithography steps are all modeled as exponentially distributed random variables. Management has provided the data on anticipated product mix and material handling requirement in Table 4, as well as the available machines and transporters (Table 5).

The current system was built five years ago; it contains only slow machines at each station and has conveyors as transporters. Now management wants to improve the performance of the system, which is measured by the long-run average cycle time. The question is which stations/transporters are worthy of investment, and whether to buy newer, faster machines that cost more, or additional slower machines that are cheaper.

To solve the problem by CSB, we model the number of fast/slow machines at each station and number of each kind of transporter as factors. All of them are discrete. The current number of each type of machine/transporter will be taken as the low setting of each factor (therefore the number of fast machines at each station and the number of AGVs is 0); the high setting of each factor is determined by the cost of each machine/transporter, as discussed in §2.2. The factors are indexed by their weight w_i, $i = 1, 2, \ldots, 20$, in ascending order to improve the efficiency of CSB (see §3.3). The factors and levels are given in Table 6.

The simulation programming of the manufacturing system is done in `simlib`, a collection of ANSI-standard C-support functions for simulation (Law and Kelton 2000). CRN is implemented by assigning each station a separate stream of random numbers. CSB with the fully sequential test is implemented in C++. For each replication, 365 days of operation are simulated with a 300 hour warm-up period to eliminate the influence of initial conditions. The performance measure is the long-run average cycle time (hours)

Table 4. Production mix and passes.

Product types	Mix% (%)	Passes required for diffusion and lithography process
A	15	20
B	35	15
C	30	10
D	20	12

Table 5. Mean processing time per cassette for each step (hours) and cost of machines ($millions).

Stations	Fast machine	Cost per unit	Slow machine	Cost per unit
CLEAN	1.5	1.38	2.5	0.83
LOAD QUARTZ	0.19	0.63	0.31	0.38
OXIDIZE	3.5	3.25	5.4	1.95
UNLOAD QUARTZ	0.19	0.63	0.31	0.38
TEST 1	0.5	1.25	1.25	0.75
COAT	0.75	1.13	1.50	0.68
STEPPER	0.85	2.25	1.8	1.35
DEVELOP	0.38	0.25	0.63	0.15
TEST 2	0.5	1.25	1.25	0.75
AGV	0.028	1.05	NA	NA
CONVEYOR	NA	NA	0.19	0.635

weighted by the percentage of different products. The quantity c^* is the price to buy one fast oxidizing machine, which equals \$3.25 million; Δ_0 is the minimum acceptable decrease in long-run cycle time that would justify a capital expenditure of \$3.25 million, and Δ_1 is the decrease in long-run cycle time that we do not want to miss if it can be achieved for \$3.25 million. The screening results are given in Table 7 with different combinations of Δ_0 and Δ_1.

To summarize, the bigger Δ_0 and Δ_1 are, the fewer factors are identified as important, and the fewer replications are required. The factors identified as important are consistent in these four cases. When Δ_0 increases to 5 and Δ_1 increases to 8, the single most important factor is determined, which is the number of AGVs in the system.

7. Conclusion

CSB is a new factor-screening method for discrete-event simulations. It combines a two-stage hypothesis-testing procedure with the SB method to control the power at each bifurcation step and Type I error for each factor under heterogeneous variance conditions. CSB is the first factor-screening procedure to provide these guarantees. Under some circumstances, a more efficient fully sequential testing procedure is available with the same error control.

It should be noted that CSB is not universally best for all factor-screening problems. When the number of factors is small or the fraction of important factors is high, CSB is not as efficient as traditional screening strategies. Also, for the tests discussed in this paper, the guarantees of performance are only true if Model (2) holds, which requires that the linear approximation is appropriate, the random errors are normally distributed, and all βs are positive. Users should be aware that CSB will miss those factors with large interactions if their main effects are not important. Fortunately, research has demonstrated that the tests are robust to moderate departures from normality (Nelson and Goldsman 2001).

Table 6. Factor description and levels (unit number).

Factor id	Factor description	Low level	High level
1	Number of slow machines in OXIDIZE	92	93
2	Number of fast machines in STEPPER	0	1
3	Number of fast machines in COAT	0	2
4	Number of slow machines in CLEAN	42	45
5	Number of fast machines in TEST 1	0	2
6	Number of fast machines in TEST 2	0	2
7	Number of slow machines in STEPPER	30	32
8	Number of slow machines in COAT	25	29
9	Number of fast machines in CLEAN	0	2
10	Number of slow machines in TEST 1	21	25
11	Number of slow machines in TEST 2	21	25
12	Number of slow machines in LOAD QUARTZ	5	13
13	Number of slow machines in UNLOAD QUARTZ	5	13
14	Number of fast machines in LOAD QUARTZ	0	5
15	Number of fast machines in UNLOAD QUARTZ	0	5
16	Number of AGVs	0	5
17	Number of slow machines in DEVELOP	10	31
18	Number of CONVEYORS	6	9
19	Number of fast machines in OXIDIZE	0	1
20	Number of fast machines in DEVELOP	0	13

Table 7. Screening results with different Δ_0 and Δ_1.

$\{\Delta_0, \Delta_1\}$ (Hours)	Important factors	Number of replications required
$\{1, 2\}$	2, 3, 5, 6, 12, 13, 15, 16, 17, 20	8,420
$\{2, 4\}$	3, 6, 12, 13, 16, 17, 20	1,439
$\{2, 5\}$	6, 12, 16	535
$\{5, 8\}$	16	289

Future research will concentrate on developing a more robust procedure which allows for interactions between factors. Another topic worth considering is how to make the procedure more adaptive to accumulated information as the screening experiment progresses.

Acknowledgments

Research for this paper was partially supported by a grant from General Motors R&D and National Science Foundation grants DMI-0217690 and DMI-0140385. The case study was based on a course project from the Department of Industrial Engineering at Arizona State University.

References

Bettonvil, B., J. P. C. Kleijnen. 1997. Searching for important factors in simulation models with many factors: Sequential bifurcation. *Eur. J. Oper. Res.* **96**(1) 180–194.

Box, G. E. P., N. R. Draper. 1987. *Empirical Model-Building and Response Surfaces*. John Wiley and Sons, New York.

Campolongo, F., J. P. C. Kleijnen, T. Andres. 2000. Screening methods. A. Saltelli, K. Chan, E. M. Scott, eds. *Sensitivity Analysis*. John Wiley and Sons, New York, 65–89.

Cheng, R. C. H. 1997. Searching for important factors: Sequential bifurcation under uncertainty. S. Andradóttir, K. J. Healy, D. H. Withers, B. L. Nelson, eds. *Proc. 1997 Winter Simulation Conf.*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 275–280. Retrieved July 6, 2003 http://www.informs-cs.org/wsc97papers/0275.PDF.

Cheng, S., C. F. J. Wu. 2001. Factor screening and response surface exploration—Rejoinder. *Statist. Sinica* **11** 553–604.

Dean, A. M., S. M. Lewis, eds. 2005. *Screening*. Springer-Verlag, New York.

Elster, C., A. Neumaier. 1995. Screening by conference designs. *Biometrika* **82**(3) 589–602.

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

Kim, S.-H. 2005. Comparison with a standard via fully sequential procedures. *ACM TOMACS*. **15**(2) 155–174.

Kleijnen, J. P. C., B. Bettonvil, F. Persson. 2006. Finding the important factors in large discrete-event simulation: Sequential bifurcation and its applications. A. Dean, S. Lewis, eds. *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*. Springer-Verlag, New York.

Law, A. M., W. D. Kelton. 2000. *Simulation Modelling and Analysis*, 3rd ed. McGraw-Hill, New York.

Lewis, S. M., A. M. Dean. 2001. Detection of interactions on large numbers of factors. *J. Roy. Statist. Soc. B* **63**(4) 633–672.

Morrice, D. J., I. R. Bardhan. 1995. A weighted least-squares approach to computer simulation factor screening. *Oper. Res.* **43**(5) 792–806.

Morris, M. D. 2006. An overview of group factor screening. A. Dean, S. Lewis, eds. *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*. Springer-Verlag, New York.

Myers, R. H., D. C. Montgomery. 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. John Wiley and Sons, New York.

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

Trocine, L., L. C. Malone. 2000. Finding important independent variables through screening designs: A comparison of methods. J. A. Joines, R. R. Barton, K. Kang, P. A. Fishwick, eds. *Proc. 2000 Winter Simulation Conf.*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 749–754. Retrieved July 6, 2003 http://www.informs-cs.org/wsc00papers/098.PDF.

Trocine, L., L. C. Malone. 2001. An overview of newer, advanced screening methods for the initial phase in an experimental design. B. A. Peters, J. S. Smith, D. J. Medeiros, M. W. Rohrer, eds. *Proc. 2001 Winter Simulation Conf.*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 169–178. Retrieved July 6, 2003 http://www.informs-cs.org/wsc01papers/020.PDF.

Wu, C. F. J., M. Hamada. 2000. *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley and Sons, New York.