Two-Stage Controlled Fractional Factorial Screening for Simulation Experiments

HONG WAN

Purdue University, 315 N. Grant Street, West Lafayette, IN 47907-2023

BRUCE E. ANKENMAN

Northwestern University, 2145 Sheridan Road, Evanston, IL 60208

Factor screening with statistical control makes sense in the context of simulation experiments that have random error, but can be run automatically on a computer and thus can accommodate a large number of replications. The discrete-event simulations common in the operations research field are well suited to controlled screening. In this paper, two methods of factor screening with control of Type I error and power are compared. The two screening methods are both robust with respect to two-factor interactions and nonconstant variance. The first method is an established sequential method called controlled sequential bifurcation for interactions (CSB-X). The second method uses a fractional factorial design in combination with a two-stage procedure for controlling power. The two-stage controlled fractional factorial (TCFF) method requires less prior information and is more efficient when the percentage of important factors is 5% or higher.

Key Words: Computer Experiments; Design of Experiments; Discrete-Event Simulation; Factor Screening; Sequential Bifurcation.

THE TERM *computer experiment* is common in the statistical literature (see Santner et al. (2003)) and refers to an experiment on a computer simulation of a physical system such as a material or a process. Typically, computer experiments are thought to be deterministic (having no random error other than rounding error). This is because many engineering simulations use finite-element or other differential equation-based simulation methods that have no random component. However, in the operations research community, the term "simulation" is most commonly used to refer to discrete-event simulations of manufacturing or service systems (see Banks et al. (2005)). These simulations are typically complex queueing systems that simulate products or customers flowing through a system. The simulation is driven by random-number generators that draw random arrival times for the products or customers and random service and failure times for the machines or stations. The input distributions for these random-number generators are based on the distributions observed for each type of machine or station in the system. Typical software packages for this type of discrete-event simulation include Witness®, Arena®, Simul8®, and Factory Explorer®. Many other software packages have been developed for more specific scenarios, for example, MedModel® for health-care simulation. Although they do have random noise like physical experiments, discrete-event simulation experiments share some attributes with computer experiments:

1. In many cases, the whole experiment can be programmed to run automatically.
2. Compared with physical experiments, it is usu-

Dr. Wan is an Assistant Professor in the School of Industrial Engineering. Her email and web addresses are hwan@purdue.edu and web.ics.purdue.edu/hwan/index.htm.

Dr. Ankenman is an Associate Professor in the Department of Industrial Engineering and Management Sciences. His email and web addresses are ankenman@northwestern.edu and www.iems.northwestern.edu/~bea

ally cheaper and easier to switch factor settings in a simulation experiment. Therefore, the design of the simulation experiment is often sequential, using a design criterion to choose the next run or to stop the experiment.

3. In many cases, replications are not as expensive as they are in physical experiments. This fact allows for the experiment size to be determined by a desired precision (or correctness) instead of a preset experimental budget.

In this article, we will compare two approaches for factor screening in the context of discrete-event simulation experiments. Factor screening refers to an experiment that is designed to sort a large set of factors into those that have an important effect on the system in question and those that do not substantially affect the system response (see Dean and Lewis (2005)). Factor screening is used for many different purposes. It is often used as a first stage of an optimization or investigation procedure intended to reduce the number of factors that must be included in more detailed investigation. However, there are many other reasons for performing factor screening, including (1) gaining a better understanding of a system by finding the factors that affect the output and, equally important, finding those factors that have a negligible effect on the output over the range of interest, (2) preparing for future adjustment or control of a system by finding a set of factors that can affect the output, and (3) simplifying the simulation model by removing portions of the model that have little effect on the output of interest.

Typically in factor screening, a hierarchical argument is used so factors with important main effects are first screened out and then additional experiments are done to determine if those important factors interact with each other or have higher order effects. Because the cost of replication is so high, the emphasis in factor screening for physical experiments has been to allow for the estimation of as many main effects as possible in as few runs as possible. This has lead to approaches such as supersaturated designs (see Lu and Xu (2004)) and group screening (see Lewis and Dean (2001)). In view of the three characteristics of discrete-event simulation experiments listed above, the screening methods that we will compare in this article will be driven not by a specific number of experimental runs but instead by a desired level of Type I error (the probability of declaring a factor important when it is not) and power (the probability of correctly declaring a factor important). Due to the statistical control that we impose on this screening experiment, we will call this *controlled screening*. Of course, given that the desired level of statistical control (power and level of Type I error) is achieved, it is desirable to reduce the number of replications and thereby reduce the total amount of computer time needed for the screening experiment. Thus, the number of replications needed to achieve the desired level of control will be the criterion for comparing the two screening methods.

In discrete-event simulations, the responses of interest are often system-performance measures like average cycle time for a product to pass through the system. For complicated systems, there are often hundreds of factors. For example, there may be many stations or machines in the manufacturing system and each station or machine may have several factors associated with it. In many cases, these factors have a known direction of effect on the response. For example, increasing the speed of a transporter will likely decrease the average cycle time if it has any effect at all. Using the assumption that the directions of the effects of the factors are known and the hierarchical assumption that linear effects are a good indicator of factor importance, Wan et al. (2005a) proposed controlled sequential bifurcation (CSB), a controlled factor-screening method for simulation experiments that uses a form of binary search called sequential bifurcation. Sequential bifurcation was originally proposed for deterministic screening by Bettonvil and Kleijnen (1997). CSB was extended to provide error control for screening main effects even in the presence of two-factor interactions and second-order effects of the factors (Wan et al. (2005b)). The new method is called CSB-X. An interesting finding in Wan et al. (2005b) is that, if interactions are not present, CSB-X is usually as efficient as CSB in identifying main effects (this depends on the variance structure at different levels of the factors). When two-factor interactions are present, CSB is unable to correctly control the identification of the main effects due to confounding with the interactions. Thus, CSB-X is recommended for controlled screening in almost all cases. The primary purpose of this paper is to compare CSB-X with another method of screening, namely, a fractional factorial design that is used as part of a two-stage method for controlling Type I error and power.

In the next section, we briefly describe the objectives of controlled screening. This structure was first introduced in Wan et al. (2005a). We then briefly de-

scribe CSB-X and introduce a version of fractional factorial screening that controls Type I error and power using a two-stage sampling procedure similar to the one used in Bishop and Dudewicz (1978) for a two-way ANOVA layout. Finally we compare the two methods under a variety of scenarios and shows the circumstances under which each method is preferred. Conclusions are then drawn.

Objective of the Controlled Screening

The purpose of this paper is to compare the performance of two methods of controlled screening. In order to compare the methods fairly, some structure on the screening problem is now established. We begin with the model. Suppose there are K factors and a simulation output response represented by Y, then we assume the following full second-order model:

$$Y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + \sum_{k=1}^{K} \sum_{m=k}^{K} \beta_{km} x_k x_m + \varepsilon, \quad (1)$$

where x_k represents the level of factor k and ε is a normally distributed error term, the variance of which is unknown and may depend on the levels of the factors, i.e., $\varepsilon \sim \mathrm{N}(0, \sigma^2(\mathbf{x}))$, where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. In this paper, we assume that the factor levels are coded on a scale from -1 to 1. Later, we will focus on Resolution IV, two-level designs that are often used in main effects screening with non-negligible interactions. In these designs, the linear main-effects estimators are not confounded by any other terms; however, many of the second-order interactions may be confounded together and the pure quadratic terms will be confounded with the estimate of β_0. Other designs could, in principle, be adapted to the proposed method, but we leave this for future research. To ease the discussion, we also assume that larger values of the response, Y, are more desirable. The case of a "smaller is better" response is easily handled by making the response negative.

The goal of the controlled screening is to provide an experimental design and analysis strategy to identify which factors have large linear main effects, the β_k's in (1). Two thresholds are established:

1. Δ_0 is the *threshold of importance* used to establish the level of Type I error. Specifically, if $|\beta_k| < \Delta_0$, then the probability of declaring factor k important should be less than α, where α is selected by the experimenter. Typically, $\alpha = 0.05$.
2. Δ_1 is the *critical threshold* used to control the power. Specifically, if $|\beta_k| \leq \Delta_1$, then the probability of declaring factor k important should be greater than γ, where γ is selected by the experimenter. Typically, $\gamma \geq 0.80$.

The scaling of the factors can be done in any way that the experimenter finds useful; however, because all factor effects are to be compared with the same thresholds, they should be comparable in some sense. One method would be to have the response be dollars of profit produced and thus all costs for changing factor levels would be incorporated into the profit calculation. Another scaling method, introduced in Wan et al. (2005a), assures that all the factor effects are on the same cost scale. For purposes of discussion, we will follow the cost-based scaling method of Wan et al. (2005a). The method assumes that the lower the factor setting, the lower the cost. The scaling is such that the cost to change each factor from the level (0) to the highest cost level (1) is a fixed cost, say c^*. Wan et al. (2005a) shows how the levels of discrete factors (e.g., the number of machines or number of operators) can be incorporated into this cost-based scaling system by introducing a weighting for each factor. The benefit of this cost-based scaling is that each linear effect, β_k, can be compared directly to the thresholds and to the other linear effects because β_k represents the expected change in the response when spending c^* to change factor k. The two thresholds now also become easily expressed concepts:

- The *threshold of importance*, Δ_0, is the minimum amount by which the response must rise in order to justify an expenditure of c^* dollars.
- The *critical threshold*, Δ_1, is an increase in the response that should not be missed if it can be achieved for an expenditure of c^* dollars.

Because the response is to be maximized, then it is natural to assume that increasing the cost of any factor will either have no effect or increase the response. Thus, in Wan et al. (2005a), it is assumed that the directions of all main effects are known and the levels of the factors are set to make all the main effects positive, i.e., $\beta_k \geq 0$, $0 < k \leq K$.

Methodology

In this section, we will first review CSB-X, the method proposed in Wan et al. (2005b) that provides controlled-factor screening in the presence of two-factor interactions. We will then present a method of controlled screening that uses a two-stage hypothesis-testing procedure in a standard fractional

factorial design. The second method will be called TCFF for two-stage controlled fractional factorial.

CSB-X, Controlled Sequential Bifurcation with Interactions

CSB-X is a controlled version of sequential bifurcation. Sequential bifurcation is a series of steps, where within each step, a group of factors is tested. Within the group, all factors are completely confounded together (as in group screening; see Lewis and Dean (2001)). Therefore, it is critical to CSB-X that the direction of all the factor effects be positive (i.e., $\beta_k \geq 0$, $0 < k \leq K$) to prevent cancellation of effects. If a group of factors is tested and the group effect (the sum of all the main effects of the factors in the group) is not found to be significantly greater than Δ_0 (the threshold of importance), then all the factors in the group are declared unimportant and are dropped from further screening. If the group effect is found to be greater than Δ_0 and there is only one factor in the group, then that factor is declared to be important. If the group effect is found to be greater than δ_0 and there is more than one factor in the group, the group is split into two subgroups that are to be tested later. This procedure is followed until all factors are classified as either important or unimportant.

Kleijnen and Bettonvil (1997) suggest various practical methods for splitting and testing the groups. Specifically, splitting groups into subgroups, the size of which is a power of 2, can improve the efficiency of the procedure. Also, if one selects testing groups that have large group effects before groups that have small group effects, the procedure can sequentially reduce the upper bound on the greatest effect that has not yet been identified. This upper bound is very useful if the procedure is stopped early. Because we are comparing the statistical control of the procedures, we plan to complete the screening procedure in all cases, and thus the sequential upper bound is not of interest. Also, we found in empirical testing that splitting groups into subgroups, the size of which is a power of 2, did not always improve efficiency and never led to increases in efficiency that substantially changed our conclusions. Thus, for better comparison to Wan et al. (2005b), we will use the procedure that simply splits each important group into two equally sized subgroups. In the case of an odd group size, the subgroup sizes differ by one. We will also follow the testing procedure of Wan et al. (2005b). To define the design points of their experiment, the design point at level k is defined as

$$x_i(k) = \begin{cases} 1, & i = 1, 2, 3, \cdots, k \\ 0, & i = k+1, k+2, \ldots, K. \end{cases}$$

Thus, "level k" has all factors with indices greater than k set to 0 on the -1 to 1 scale and all factors with indices less than or equal to k, set to 1. Similarly, a design point at "level $-k$" is defined as

$$x_i(-k) = \begin{cases} -1, & i = 1, 2, 3, \cdots, k \\ 0, & i = k+1, k+2, \ldots, K. \end{cases}$$

To test a group of factors, say factors with indices in the set $\{k_1 + 1, k_1 + 2, \ldots, k_2\}$, four design points are used: level k_1, level $-k_1$, level k_2, and level $-k_2$. To follow the notation in Wan et al. (2005b), let $Z_\ell(k)$ represent the observed simulation output for the ℓth replication at level k. The procedure ensures that all four design points have the same number of replications and the test statistic, $D_\ell(k_1, k_2)$, is calculated as

$$D_\ell(k_1, k_2) = \frac{[Z_\ell(k_2) - Z_\ell(-k_2)] - [Z_\ell(k_1) - Z_\ell(-k_1)]}{2}.$$

They show that, given (1), the expected value of $D_\ell(k_1, k_2)$ is the sum of the factors' main effects from the group $\{k_1 + 1, k_1 + 2, \ldots, k_2\}$, i.e.,

$$E[D_\ell(k_1, k_2)] = \sum_{k=k_1+1}^{k_2} \beta_k.$$

The group is declared important if the null hypothesis, H_0: $\sum_{k=k_1+1}^{k_2} \beta_k \leq \Delta_0$, is rejected at level α. The group is declared unimportant if the null hypothesis is not rejected, provided that the test used has power γ for the alternative hypothesis, H_A: $\sum_{k=k_1+1}^{k_2} \beta_k \geq \Delta_1$. To guarantee this statistical control, Wan et al. (2005b) provide a fully sequential test that first takes an initial number, n_0, of replications at each of the four design points and then adds replications one at a time (to all four design points) until the group effect can be declared either important or unimportant. Even under nonconstant variance conditions as in (1), CSB-X with the fully sequential test guarantees that the probability of Type I error is α for each factor individually and the power is at least γ for each step of the algorithm.

TCFF, Two-Stage Controlled Fractional Factorial

The TCFF method begins by selecting a fractional factorial design that is capable of achieving the desired resolution for the number of factors being

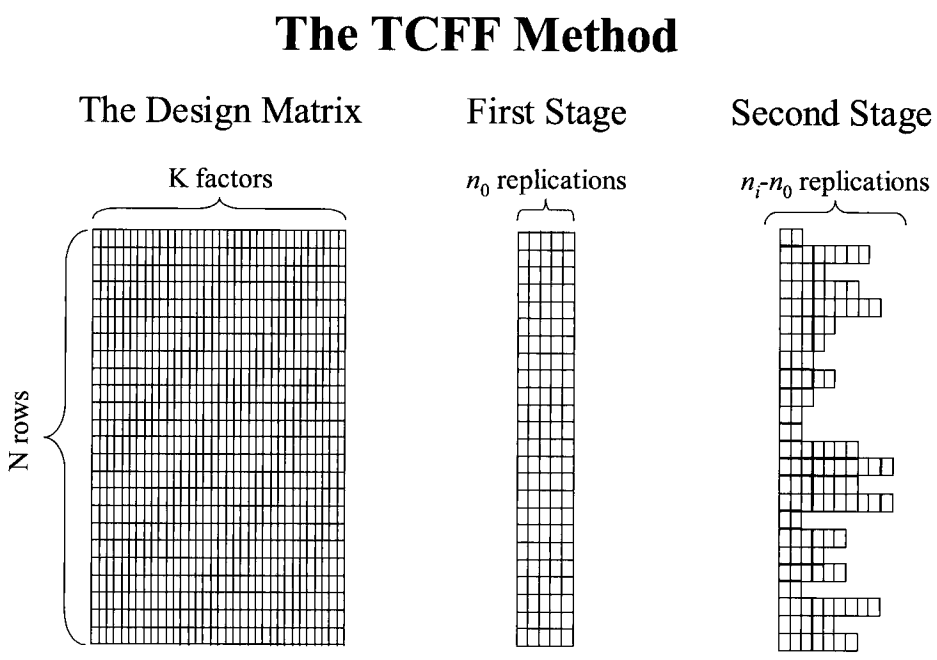

FIGURE 1. A Graphical View of the TCFF Method.

screened. Recall that a Resolution III design can estimate all main effects independent of each other but does confound main effects with certain two-factor interactions (similar to CSB). A Resolution IV design also allows for independent estimation of main effects and, in addition, removes any bias due to two-factor interactions between the factors. In order to match the capabilities of CSB-X, we will use Resolution IV designs for the comparison in this paper.

Many books (e.g., Wu and Hamada (2000)) and software packages provide recommended fractional factorial designs for various resolutions and various values of K, the number of factors to be screened. If a recommended design cannot be easily found for a particular value of K, a Resolution IV design can always be easily constructed by taking an orthogonal array large enough to accommodate K two-level factors and then folding it over according to the procedure described on page 172 of Wu and Hamada (2000). Because the foldover procedure doubles the number of runs in the orthogonal array, this procedure produces a Resolution IV design for K factors that has at least $2 \times (K+1)$ design points. A very extensive list of orthogonal arrays for different values of K has recently been published by Kuhfield and Tobias (2005).

The method described below for controlling the power does not require any particular resolution design and in fact can be used with higher resolution designs if one wants to screen for factors with important two-factor interactions or higher order effects. Of course higher resolution designs will require more design points and potentially many more observations.

Once the fractional factorial design (with N design points) is selected, the next step is to control the power and Type I error of the main-effect estimates. To accomplish this, we adapt a two-stage method developed to control power for the two-way ANOVA layout by Bishop and Dudewicz (1978). The first stage of the method requires an initial number, say n_0, of replications of the fractional factorial design, i.e., n_0 observations at each design point. Because the fractional factorial is likely to be quite large, we recommend that n_0 be relatively small, such as 3 or 5. The second stage of the method adds replications to certain rows of the factorial design to get better estimates of the mean response of those rows for which the responses have high variance. We denote the total number of replications for row i after both stages as n_i. Thus, the number of additional replications taken for row i in the second stage is $n_i - n_0$. The TCFF procedure is represented graphically in Figure 1. The keys to the control of Type I error and power for this method are: (1) the calculation for n_i and (2) careful selection of the estimators for each factor effect.

Before describing the calculations for n_i, a brief overview of the analysis will be helpful. Once all the replications for stage 2 are made for each row, a single pseudo-observation, called $\tilde{Y}_i$, will be calculated for each row. Each $\tilde{Y}_i$ will be a weighted average of all the observations from row i. Given the normal error structure shown in Equation (1), careful selection of the number of runs in each row and the weights in the weighted averages will result in each of these $\tilde{Y}_i$'s having a distribution which is proportional to a non-central t-distribution with $n_0 - 1$ degrees of freedom. Once the pseudo-observations are calculated for each row, the design is treated like an unreplicated fractional factorial design and the main effects of each factor are calculated in the standard way (either with contrasts as described on page 228 of Montgomery, 2001 or through regression analysis, see page 113 of Wu and Hamada, 2000). Since each of the $\tilde{Y}_i$'s follows a t-distribution, the null distribution of each effect estimate can be found through Monte Carlo simulation or through a normal approximation. Thus, an α-level test can be used to test the hypothesis: H_0: $|\beta_k| \leq \Delta_0$ or H_1: $|\beta_k| > \Delta_0$. The calculation of the number of replications in the second stage of the method ensures that if the effect of factor k is greater than Δ_1, the probability that factor k will be declared important is greater than γ.

The steps for determining n_i are now given (The formulas are adapted from Bishop and Dudewicz, 1978.):

1. Let the CDF of the distribution of the average of N independent standard t-distributed random variables each with $n_0 - 1$ degrees of freedom be called $\bar{t}(N, n_0 - 1, x)$, where x is the argument of the CDF. Let c_0 be the $1-\alpha$ quantile of this distribution and c_1 be the $1 - \gamma$ quantile of this distribution. These quantiles can be easily obtained through simulation or through a normal approximation (see the Appendix). Values for c_0 and c_1 for some common values of n_0 and N are given in the tables in the Appendix.
2. Let $z = [(\Delta_0 - \Delta_1)/(c_0 - c_1)]^2$ and let s_i be the sample standard deviation from the n_0 replicates of row i from the first stage. The total number of replications needed for row i after the second stage of the method is then

$$n_i = \max\left(n_0 + 1, \left\lfloor \frac{s_i^2}{z} \right\rfloor + 1\right),$$

where $\lfloor x \rfloor$ denotes the greatest integer less than x.

Thus, in the second stage, $n_i - n_0$ new observations are taken for each row i.

After all observations are taken, each factor effect, β_k, is tested as follows:

1. For each row in the fractional factorial design, a weight for each replication must be calculated. The jth replication in the ith row will be called y_{ij} and the associated weight will be called a_{ij}.
2. A preliminary value called b_i is first calculated for each row as

$$b_i = \frac{1}{n_i}\left[1 + \sqrt{\frac{n_0(n_i z - s_i^2)}{(n_i - n_0)s_i^2}}\right],$$

where n_0 is the number of initial replications in the first stage, n_i is the total number of replications of row i after the second stage, $z = [(\Delta_0 - \Delta_1)/(c_0 - c_1)]^2$, and s_i is the sample standard deviation from the initial n_0 replicates of row i from the first stage.
3. The weight for the jth replication in the ith row is

$$a_{ij} = \begin{cases} \dfrac{1 - (n_i - n_0)b_i}{n_0} & \text{for } j = 1, 2, \ldots, n_0, \\ b_i & \text{for } j = n_0 + 1, n_0 + 2, \ldots, n_i. \end{cases}$$

4. These weights are then used to calculate a single weighted average response for each row, called $\tilde{Y}_i$, where $\tilde{Y}_i = \sum_{j=1}^{n_i} a_{ij} y_{ij}$. Given Equation (1), each $\tilde{Y}_i$ has a scaled non-central t-distribution such that

$$\tilde{Y}_i \sim \sqrt{z}t_i + \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{k=1}^{K}\sum_{m=k}^{K} \beta_{km} x_{ik} x_{im}, \tag{2}$$

where each t_i is an independent random variable that has a central t-distribution with $n_0 - 1$ degrees of freedom.
5. The design can now be viewed as an unreplicated fractional factorial design with observations $\tilde{Y}_i$. Let the level of the kth factor in the ith row of the design matrix be x_{ik}. The estimator for the coefficient of the main effect for the kth factor is

$$\hat{\beta} + \frac{1}{N}\sum_{i=1}^{N} x_{ik}\tilde{Y}_i. \tag{3}$$

6. Each x_{ik} in the fractional factorial design will be either 1 or -1, and each column, $\mathbf{x}_k$, is orthogonal to the mean, all the other main effects, and all two-factor interaction columns.

Since the t-distribution is symmetric (i.e., $-t_i$ has the same distribution as t_i), it follows from (2) and (3) that

$$\hat{\beta}_k \sim \beta_k + \sqrt{z}\sum_{i=1}^{N}\frac{t_i}{N}, \qquad (4)$$

where each t_i is an independent random variable that has a central t-distribution with n_0-1 degrees of freedom.

7. We now test the hypothesis: H_0: $|\beta_k| \leq \Delta_0$ or H_1: $|\beta_k| > \Delta_0$. The null hypothesis is rejected only if $|\hat{\beta}_k| > \Delta_0 + c_0\sqrt{z}$ since it follows from (4) that

$$\Pr\left[|\hat{\beta}_k| > \Delta_0 + c_0\sqrt{z}\,\middle|\,|\beta_k| < \Delta_0\right] < \alpha,$$

where c_0 is the $1-\alpha$ quantile of the distribution with CDF $\bar{t}(N, n_0-1, x)$.

Bishop and Dudewicz (1978) use a theorem from Stein (1945) to show that for a two-way ANOVA layout, this procedure provides exact control of Type I error even if the variance changes across treatment combinations. They also extend the results to a multi-way ANOVA in Bishop and Dudewicz (1981). Since the fractional factorial design can be viewed as a multi-way ANOVA with various main effects fully confounded with high order interactions whose effects are assumed to be zero, the procedure guarantees exact control of Type I error given the orthogonal design and the model in Equation (1).

Power control is achieved by choosing the value for z. The null hypothesis is rejected only if $|\hat{\beta}_k| > \Delta_0 + c_0\sqrt{z}$. By Equation (4), $\Pr[|\hat{\beta}_k| > \Delta_1 + c_1\sqrt{z} \mid |\beta_k| \geq \Delta_1] \geq \gamma$. To control power, z must then be set such that $\Delta_1 + c_1\sqrt{z} = \Delta_0 + c_0\sqrt{z}$, so that, if $\beta \geq \Delta_1$, the probability of rejecting the null hypothesis is greater than γ (i.e., $\Pr[|\hat{\beta}_k| > \Delta_0 + c_0\sqrt{z} \mid |\beta_k| \geq \Delta_1] \geq \gamma$).

Therefore,

$$z = \left(\frac{\Delta_1 - \Delta_0}{c_0 - c_1}\right)^2.$$

This concludes the methodology for TCFF. In the next section, a simple numerical example with a small number of factors is provided to help clarify the computations.

Numerical Example

Suppose that a small manufacturing system has two stations and, at each station, there are three factors: the number of machines (coded M1 and M2), the number of operators (coded O1 and O2), and the frequency of preventative maintenance (coded F1 and F2). The goal is to use a 16-run Resolution IV

TABLE 1. Data from the First Stage of the Numerical Example

Design matrix							First-stage replications				
Row	M1	M2	O1	O2	F1	F2	y_{i1}	y_{i2}	y_{i3}	y_{i4}	S_i
1	−1	−1	−1	−1	−1	−1	10035	9110	8995	8758	560
2	1	−1	−1	−1	1	−1	8036	7462	8105	9866	1040
3	−1	1	−1	−1	1	1	8580	8838	8814	10228	751
4	1	1	−1	−1	−1	1	12744	14731	13924	12051	1196
5	−1	−1	1	−1	1	1	10168	10976	11008	9799	602
6	1	−1	1	−1	−1	1	12305	11929	10099	10961	993
7	−1	1	1	−1	−1	−1	9342	8551	8650	8392	419
8	1	1	1	−1	1	−1	9073	9735	12433	10260	1455
9	−1	−1	−1	1	−1	1	9180	8109	10432	12130	1729
10	1	−1	−1	1	1	1	11469	11415	12411	10945	614
11	−1	1	−1	1	1	−1	8052	8317	8392	8268	146
12	1	1	−1	1	−1	−1	11295	9293	9248	8981	1069
13	−1	−1	1	1	1	−1	9040	7253	9001	8179	843
14	1	−1	1	1	−1	−1	8710	9359	9029	9820	475
15	−1	1	1	1	−1	1	8877	11124	9329	9755	970
16	1	1	1	1	1	1	12710	11700	11371	15765	2002

TABLE 2. Data from the Second Stage of the Numerical Example

	Second-stage replications and computations									
Row	y_{i5}	y_{i6}	y_{i7}	y_{i8}	y_{i9}	y_{i10}	y_{i11}	y_{i12}	b_i	$\tilde{Y}_i$
1	7386								1.058	7279
2	8470								0.516	8420
3	8139								0.781	8352
4	14696								0.391	13884
5	7781								0.985	7821
6	9954								0.553	10566
7	8437								1.399	8318
8	8997	8930	10503						0.209	9812
9	9838	9769	8724	10936	10204				0.135	9917
10	10242								0.965	10289
11	8054								3.808	7483
12	11843								0.493	10758
13	9810								0.685	9356
14	9872								1.243	10028
15	10526								0.572	10203
16	11563	17353	12074	10232	13121	8399	9980	14789	0.097	12347

fractional factorial to screen for factors that have a large effect on the daily throughput of the system. Table 1 shows the design matrix and the first-stage observations. We have chosen $\Delta_0 = 300$, $\Delta_1 = 1100$, $\alpha = 0.05$, $\gamma = 0.95$, and $n_0 = 4$.

In order to calculate n_i for each row, we begin by calculating z. We can approximate c_0 and c_1, the $1 - \alpha$ and $1 - \gamma$ quantiles of the $\bar{t}$-distribution, by applying the central limit theorem and using a normal approximation. Because the variance of a t-distribution with 3 degrees of freedom is 3, the variance of the average of 16 independent t-distributed random variables should be approximately normal with variance 3/16. Thus, $c_0 \approx \sqrt{3/16}\phi^{-1}(0.95) = 0.7122$ and $c_1 \approx \sqrt{3/16}\phi^{-1}(0.05) = 0.7122$, where $\phi^{-1}(x)$ is the inverse function for the standard normal distribution. Using 10,000 Monte Carlo simulations, we found values of $c_0 = 0.675$ and $c_1 = -0.675$ (see Appendix Table A2). We will use the Monte Carlo simulation results to calculate z as follows: $z = [(\Delta_1 - \Delta_0)/(c_0 - c_1)]^2 = 351{,}166$. The calculated values for n_i are (5, 5, 5, 5, 5, 5, 5, 7, 9, 5, 5, 5, 5, 5, 5, 12) for rows 1–16, respectively. The $n_i - n_0$ observations for the second stage, the b_i's, and $\tilde{Y}_i$'s are given in Table 2. The estimates for the factor effect coefficients are shown in Table 3. Because $\Delta_0 + c_0\sqrt{z} = 700$, then we see that M1 and F2 have important effects.

Comparison of Methods

With normally distributed error, both CSB-X and TCFF have guaranteed performance for Type I error and power, even when there is unequal variance and two-factor interactions present in the model. Thus, the primary measure for comparison between the two methods is the number of replications that it takes to gain the required performance. We set up 11 scenarios under which to test the two methods. Each

TABLE 3. The Estimates for the Coefficients for the Main Effect of Each Factor

Coefficient	Mean	M1	M2	O1	O2	F1	F2
Estimate	9677	1086	468	129	370	−442	745

of these scenarios is used for 200 and 500 factors with 1%, 5%, and 10% of the factors important. The TCFF method uses a 512-run, Resolution IV design for the 200-factor cases and a 1024-run, Resolution IV design for the 500-factor cases. TCFF uses $n_0 = 3$ replications for the first stage. CSB-X uses $n_0 = 5$.

In each scenario, the size of the important effects is 5 and the size of unimportant effects is 0. We set Δ_0 to 2 and Δ_1 to 4 for all scenarios, and the error was normally distributed. The first three scenarios have equal variance. In the next five scenarios, each of the important factors also has a dispersion effect, which means that, when one of those factors changes level, the variability of the error either increases or decreases. Half the important factors increase the variability and the other half decrease the variability. Each dispersion effect increases or decreases the standard deviation of the error by 20% in the cases with 200 factors and by 8% in the cases with 500 factors. In the final three scenarios, the standard deviation of the error is proportional to the expected value of the response. The constant of proportionality was 0.1 and 0.04 for the 200 and 500 factor cases, respectively. These are very important scenarios because, for both simulation experiments and physical experiments, it is very common that variability and average response value are related. Each of the 11 scenarios is described below.

1. The *important effects are clustered* together at the beginning of the factor set, and the standard deviation for the random *error is set to 3* for all observations.
2. The *important effects are distributed* at regular intervals throughout the factor set, and the standard deviation for the random *error is set to 3* for all observations.
3. The *important effects are randomly distributed* throughout the factor set, and the standard deviation for the random *error is set to 3* for all observations.
4. The *important effects are clustered* together at the beginning of the factor set. Each of the important factors also has a dispersion effect. In this scenario, all the *positive dispersion effects are clustered* at the beginning of the factor set.
5. The *important effects are clustered* together at the beginning of the factor set. Each of the important factors also has a dispersion effect. In this scenario, the *positive dispersion effects are distributed* throughout the set of important factors.
6. The *important effects are distributed* throughout the factor set. Each of the important factors also has a dispersion effect. In this scenario, all the *positive dispersion effects are clustered* at the beginning of the set of important factors.
7. The *important effects are distributed* throughout the factor set. Each of the important factors also has a dispersion effect. In this scenario, the *positive dispersion effects are distributed* throughout the set of important factors.
8. The *important effects are randomly distributed* throughout the factor set. Each of the important factors also has a *dispersion effect, which is randomly chosen to be positive or negative.*
9. The *important effects are clustered* together at the beginning of the factor set, and the standard deviation for the random *error is proportional to the average response.*
10. The *important effects are distributed* at regular intervals throughout the factor set, and the standard deviation for the random *error is proportional to the average response.*
11. The *important effects are randomly distributed* throughout the factor set, and the standard deviation for the random *error is proportional to the average response.*

There were 10 macroreplications of each scenario for 200 and 500 factors and with 1%, 5%, and 10% of the factors being important. Because both methods eliminate the effects of two-factor interactions, interactions were randomly generated for each simulation and the two methods were given the same interaction matrix for each of the 10 macroreplications under each scenario. The interactions were generated from a $N(0, 2)$ distribution and randomly assigned between factors. Factors with important effects had a 64% chance of interaction with other important factors and a 16% chance of interaction with factors with unimportant effects. Each pair of unimportant factors had only a 4% chance of interaction. Surprisingly, using the same interaction matrix did not seem to induce much correlation between the macroreplications. The average and standard deviation of the number of replications across the 10 macroreplications for each method for each scenario are reported in Tables 4 and 5. The average number of replications is plotted across the scenarios in Figures 2–7.

It is clear from Figures 2 and 5 that CSB-X requires fewer replications when only 1% of the factors

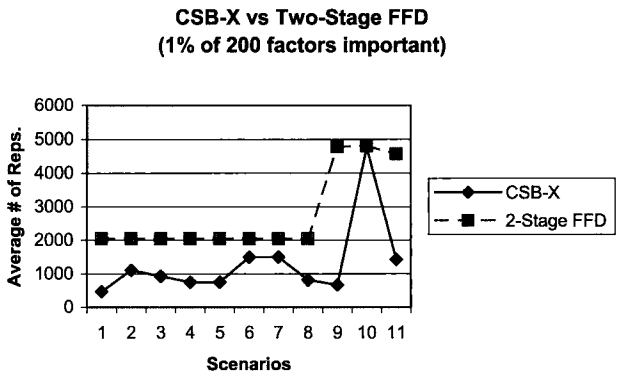

FIGURE 2. Average Number of Runs with 1% of 200 Factors Important.

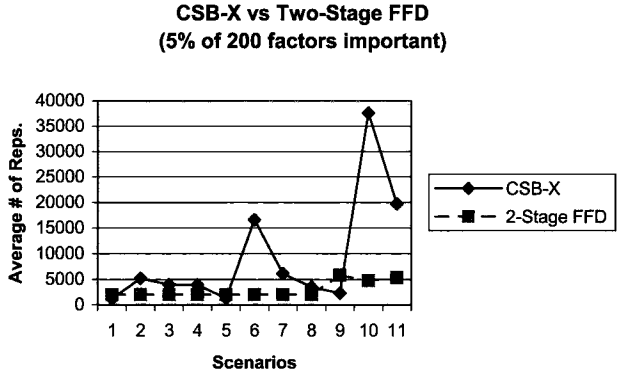

FIGURE 3. Average Number of Runs with 5% of 200 Factors Important.

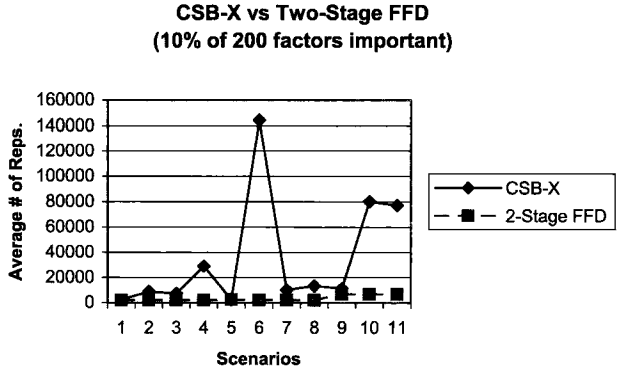

FIGURE 4. Average Number of Runs with 10% of 200 Factors Important.

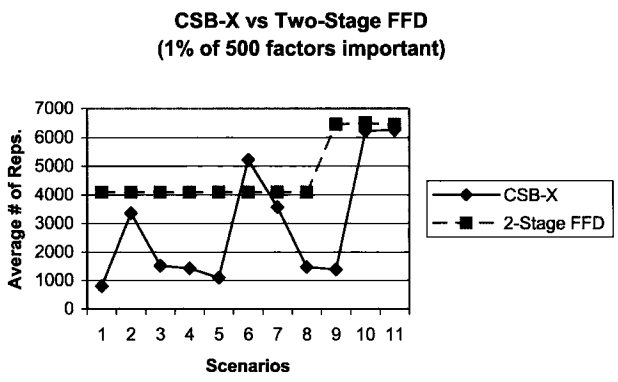

FIGURE 5. Average Number of Runs with 1% of 500 Factors Important.

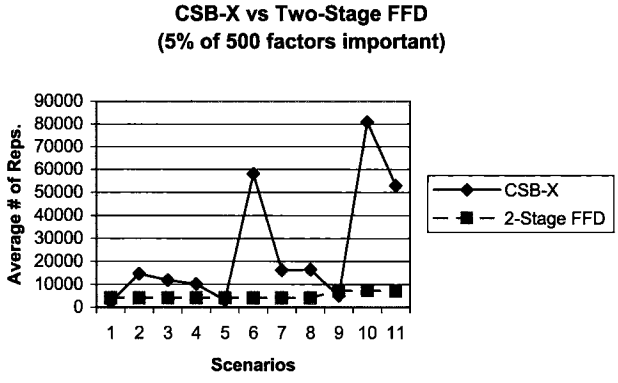

FIGURE 6. Average Number of Runs with 5% of 500 Factors Important.

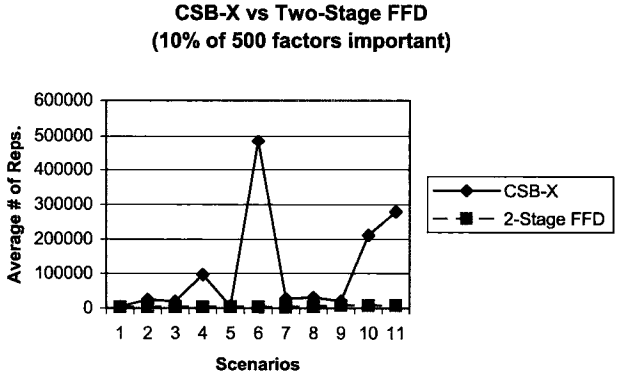

FIGURE 7. Average Number of Runs with 10% of 500 Factors Important.

TABLE 4. Simulation Results for CSB-X and TCFF with 200 Factors

Scenario	Number of factors	Number of important factors	Significant effects order	Variance model	CSB-X Average number of reps	CSB-X Standard deviation of reps	TCFF Anerage number of reps	TCFF Standard deviation of reps
1	200	2	Clustered	Equal	474	109	2048	0
2	200	2	Distributed	Equal	1108	259	2048	0
3	200	2	Random	Equal	921	181	2048	0
4	200	2	Clustered	Clust. disp. eff.	749	259	2048	0
5	200	2	Clustered	Dist. disp. eff.	749	259	2048	0
6	200	2	Distributed	Clust. disp. eff.	1500	516	2048	0
7	200	2	Distributed	Dist. disp. eff.	1500	516	2048	0
8	200	2	Random	Disp. eff.	815	135	2048	0
9	200	2	Clustered	Prop. to mean	673	586	4777	279
10	200	2	Distributed	Prop. to mean	4768	5412	4801	506
11	200	2	Random	Prop. to mean	1430	1093	4563	397
1	200	10	Clustered	Equal	1231	373	2048	0
2	200	10	Distributed	Equal	5207	979	2049	1
3	200	10	Random	Equal	4016	963	2048	0
4	200	10	Clustered	Clust. disp. eff.	3938	1603	2049	1
5	200	10	Clustered	Dist. disp. eff.	1337	528	2053	3
6	200	10	Distributed	Clust. disp. eff.	16627	3647	2049	1
7	200	10	Distributed	Dist. disp. eff.	6181	1198	2049	2
8	200	10	Random	Disp. eff.	3542	613	2048	0
9	200	10	Clustered	Prop. to mean	2305	850	5845	561
10	200	10	Distributed	Prop. to mean	37643	41773	4801	506
11	200	10	Random	Prop. to mean	19748	19692	5386	522
1	200	20	Clustered	Equal	2293	805	2048	1
2	200	20	Distributed	Equal	8996	1662	2048	0
3	200	20	Random	Equal	7290	875	2048	0
4	200	20	Clustered	Clust. disp. eff.	28962	15887	2140	33
5	200	20	Clustered	Dist. disp. eff.	2465	671	2589	179
6	200	20	Distributed	Clust. disp. eff.	144597	60070	2112	49
7	200	20	Distributed	Dist. disp. eff.	10189	1741	2227	120
8	200	20	Random	Disp. eff.	13347	2372	2052	4
9	200	20	Clustered	Prop. to mean	11176	3065	6958	1007
10	200	20	Distributed	Prop. to mean	80121	47339	6829	873
11	200	20	Random	Prop. to mean	77146	59144	6923	543

are important. In most of these cases, the TCFF method takes the minimum number of runs, $4 \times 1024 = 4096$ runs. Figures 3, 4, 6, and 7 show that the TCFF method is better on average than CSB-X when the percentage of important factors is greater than 5%.

Tables 4 and 5 show that the standard deviation across the macroreplications is always lower for TCFF and that it can be quite high for CSB-X, especially in the scenarios where the standard deviation of the error is proportional to the average response (scenarios 9, 10, and 11). Also, CSB-X is sensitive

TABLE 5. Simulation Results for CSB-X and TCFF with 500 Factors

Scenario	Number of factors	Number of important factors	Significant effects order	Variance model	CSB-X Average number of reps	CSB-X Standard deviation of reps	TCFF Average number of reps	TCFF Standard deviation of reps
1	500	5	Clustered	Equal	803	304	4096	0
2	500	5	Distributed	Equal	3368	746	4096	0
3	500	5	Random	Equal	1516	413	4096	0
4	500	5	Clustered	Clust. disp. eff.	1424	673	4096	0
5	500	5	Clustered	Dist. disp. eff.	1096	442	4096	0
6	500	5	Distributed	Clust. disp. eff.	5233	1348	4096	0
7	500	5	Distributed	Dist. disp. eff.	3567	803	4096	0
8	500	5	Random	disp. eff.	1474	334	4096	0
9	500	5	Clustered	Prop. to mean	1389	1954	6451	275
10	500	5	Distributed	Prop. to mean	6221	3305	6500	295
11	500	5	Random	Prop. to mean	6259	9382	6435	263
1	500	25	Clustered	Equal	2388	571	4096	0
2	500	25	Distributed	Equal	14713	3018	4096	0
3	500	25	Random	Equal	11837	2640	4096	0
4	500	25	Clustered	Clust. disp. eff.	10262	1025	4096	0
5	500	25	Clustered	Dist. disp. eff.	3089	588	4097	3
6	500	25	Distributed	Clust. disp. eff.	58150	14558	4097	1
7	500	25	Distributed	Dist. disp. eff.	16192	3206	4096	1
8	500	25	Random	Disp. eff.	16499	2755	4106	9
9	500	25	Clustered	Prop. to mean	5039	2055	7278	323
10	500	25	Distributed	Prop. to mean	80923	52420	7397	606
11	500	25	Random	Prop. to mean	53043	32594	7204	464
1	500	50	Clustered	Equal	4723	876	4096	0
2	500	50	Distributed	Equal	25453	4587	4096	0
3	500	50	Random	Equal	20167	2345	4096	0
4	500	50	Clustered	Clust. disp. eff.	97447	20785	4126	17
5	500	50	Clustered	Dist. disp. eff.	5491	920	4562	217
6	500	50	Distributed	Clust. disp. eff.	484380	116160	4168	53
7	500	50	Distributed	Dist. disp. eff.	27732	4909	4195	33
8	500	50	Random	Disp. eff.	31145	7961	4525	121
9	500	50	Clustered	Prop. to mean	20454	5971	8313	395
10	500	50	Distributed	Prop. to mean	211134	143867	8532	553
11	500	50	Random	Prop. to mean	279118	245735	8639	551

to the order in which the important effects appear, where TCFF is not. Naturally, CSB-X does very well when the important effects are clustered, but performs worse when they are randomly placed or distributed regularly. Finally, recall that CSB-X requires the direction of the factor effects to be known, where TCFF does not.

Conclusion

Two methods, CSB-X and TCFF, for screening of a large number of factors with statistical control of the Type I error and the power have been reviewed and compared. If there are very few important factors (less than 5%) and there is substantial prior informa-

tion (known direction of factor effects and some prior information that allows the important effects to be clustered), CSB-X can have substantial gains in efficiency over the TCFF approach; however, it can have highly variable results depending on the quality of the prior information and the extent to which the error variance changes across observations. The TCFF method requires little prior information but does involve a substantial initial investment of replications. Once this initial investment is made, the method is relatively stable and performs better than CSB-X when the percentage of important factors increases above 5%.

Appendix
Critical Values for the $\bar{t}$ Distribution

In this appendix, we show how to compute values for c_0 and c_1 through Monte Carlo simulation or with a normal approximation. Let the CDF of the distribution of the average of N independent standard t-distributed random variables each with $n_0 - 1$ degrees of freedom be called $\bar{t}(N, n_0 - 1, x)$, where x is the argument of the CDF. Let c_0 be the $1 - \alpha$ quantile of this distribution and c_1 be the $1-\alpha$ quantile of this distribution. Values for c_0 and c_1 for some common values of α, γ, n_0, and N are given in Tables A1, A2, and A3.

TABLE A1. The Critical Values, c_0, for $\alpha = 0.01$. These are also the values, $-c_1$, for $\gamma = 0.99$. Normal approximation results are in parentheses

	$N = 8$	$N = 16$	$N = 32$
$n_0 = 3$	2.626	1.885	1.426
	(n/a)	(n/a)	(n/a)
$n_0 = 4$	1.523	1.098	0.758
	(1.425)	(1.007)	(0.712)
$n_0 = 5$	1.200	0.848	0.603
	(1.163)	(0.822)	(0.582)
$n_0 = 6$	1.103	0.737	0.536
	(1.062)	(0.751)	(0.531)
$n_0 = 7$	1.017	0.726	0.517
	(1.007)	(0.712)	(0.504)
$n_0 = 8$	1.001	0.712	0.507
	(0.973)	(0.688)	(0.487)
$n_0 = 9$	0.953	0.668	0.483
	(0.950)	(0.672)	(0.475)
$n_0 = 10$	0.965	0.676	0.473
	(0.933)	(0.659)	(0.466)

TABLE A2. The Critical Values, c_0, for $\alpha = 0.05$. These are also the values, $-c_1$, for $\gamma = 0.95$. Normal approximation results are in parentheses

	$N = 8$	$N = 16$	$N = 32$
$n_0 = 3$	1.291	1.034	0.742
	(n/a)	(n/a)	(n/a)
$n_0 = 4$	0.930	0.675	0.494
	(1.007)	(0.712)	(0.504)
$n_0 = 5$	0.802	0.571	0.411
	(0.822)	(0.582)	(0.411)
$n_0 = 6$	0.759	0.523	0.378
	(0.751)	(0.531)	(0.375)
$n_0 = 7$	0.718	0.508	0.365
	(1.007)	(0.504)	(0.356)
$n_0 = 8$	0.692	0.481	0.346
	(0.688)	(0.487)	(0.344)
$n_0 = 9$	0.674	0.473	0.336
	(0.672)	(0.475)	(0.336)
$n_0 = 10$	0.665	0.465	0.330
	(0.659)	(0.466)	(0.330)

TABLE A3. The Critical Values, c_0, for $\alpha = 0.10$. These are also the values, $-c_1$, for $\gamma = 0.90$. Normal approximation results are in parentheses

	$N = 8$	$N = 16$	$N = 32$
$n_0 = 3$	0.912	0.734	0.543
	(n/a)	(n/a)	(n/a)
$n_0 = 4$	0.686	0.510	0.371
	(0.785)	(0.555)	(0.392)
$n_0 = 5$	0.606	0.444	0.319
	(0.641)	(0.543)	(0.320)
$n_0 = 6$	0.587	0.404	0.292
	(0.585)	(0.414)	(0.292)
$n_0 = 7$	0.549	0.396	0.279
	(0.555)	(0.392)	(0.277)
$n_0 = 8$	0.537	0.380	0.273
	(0.536)	(0.379)	(0.268)
$n_0 = 9$	0.518	0.361	0.263
	(0.523)	(0.370)	(0.262)
$n_0 = 10$	0.520	0.360	0.259
	(0.514)	(0.363)	(0.257)

Monte Carlo Simulation to Determine c_0 and c_1 for $\bar{t}(N, n_0 - 1, x)$

1. Generate a set of N random variables, t_i for all $i \in \{1, \ldots, N\}$, from a central t-distribution with $n_0 - 1$ degrees of freedom, and compute the average, called $\bar{t}_j = (1/N) \sum_{i=1}^{N} t_i$.
2. Choose a large number, M, and repeat step 1 M times to obtain the average of each set, $\bar{t}_j$, for all $j \in \{1, \ldots, M\}$.
3. Rank order the $\bar{t}_j$'s so that $\bar{t}_{(j)} > \bar{t}_{(k)}$, for all $j > k$, to obtain $\bar{t}_{(j)}$ for all $j \in \{1, \ldots, M\}$.
4. Let $j_0 = Round[(1-\alpha)M]$ and $j_1 = Round[(1-\gamma)M]$; then $c_0 = \bar{t}_{(j_0)}$ and $c_1 = \bar{t}_{(j_1)}$, where $Rounds[x]$ rounds x to the nearest integer.

Normal Approximation to Determine c_0 and c_1 for $\bar{t}(N, n_0 - 1, x)$

The variance of a t-distribution with v degrees of freedom is $v/(v-2)$ for $v > 2$. If N is relatively large, then by the central limit theorem, the average of N independent t-distributed random variables is approximately normal with variance $v/N(v-2)$. Thus,

$$c_0 \approx \sqrt{\frac{v}{N(v-2)}}\phi^{-1}(1-\alpha)$$

and

$$c_1 \approx \sqrt{\frac{v}{N(v-2)}}\phi^{-1}(1-\gamma),$$

where $\phi^{-1}(x)$ is the inverse function for the standard normal distribution.

In Tables A1, A2, and A3, critical values c_0 and c_1 calculated by Monte Carlo simulation (and the normal approximation) are presented for cases where the normal approximation is not very accurate. For $N > 32$, the normal approximation is reasonably accurate and thus tables are not necessary. Notice that, if $(1-\gamma) = \alpha$, then $c_1 = -c_0$. The normal approximation cannot be used when $n_0 = 3$ because there is no closed form for the variance of the t-distribution with 2 degrees of freedom.

Acknowledgment

Dr. Ankenman's work was partially supported by EPSRC grant GR/S91673 and was undertaken while he was visiting Southampton Statistical Sciences Research Institute at the University of Southampton in the UK. Dr. Wan's research was partially supported by Purdue Research Foundation (award no. 6904062). We would also like to thank Russell Cheng, Sue Lewis, Barry Nelson, and an anonymous referee for many helpful discussions and comments.

References

BANKS, J.; CARSON II, J. S.; NELSON, B. L.; and NICOL, D. M. (2005). *Discrete-Event System Simulation*, 4th ed. Prentice Hall, NJ.

BETTONVIL, B. and KLEIJNEN, J. P. C. (1997). "Searching for Important Factors in Simulation Models with Many Factors: Sequential Bifurcation". *European Journal of Operational Research* 96, pp. 180–194.

BISHOP, T. A. and DUDEWICZ, E. J. (1978). "Exact Analysis of Variance with Unequal Variances—Test Procedures and Tables". *Technometrics* 20, pp. 419–430.

BISHOP, T. A. and DUDEWICZ, E. J. (1981). "Heteroscedastic ANOVA". *Sankhya* 20, pp. 40–57.

DEAN, A. M. and LEWIS, S. M., (eds.) (2005). *Screening: Methods for Experimentation in Industry*. Springer-Verlag, New York, NY.

KUHFELD, W. F. and TOBIAS, R. D. (2005). "Large Factorial Designs for Product Engineering and Marketing Research Applications". *Technometrics* 47, pp. 132–141.

LEWIS, S. M. and DEAN, A. M. (2001). "Detection of Interactions in Experiments with Large Numbers of Factors (with Discussion)". *Journal of the Royal Statistics Society, Series B* 63, pp. 633–672.

LU, X. and WU, X. (2004). "A Strategy of Searching for Active Factors in Supersaturated Screening Experiments". *Journal of Quality Technology* 36, pp. 392–399.

MONTGOMERY, D. C. (2001). *The Design and Analysis of Experiments*, 5th ed. Wiley, New York, NY.

SANTNER, T. J.; WILLIAMS, B.; and NOTZ, W. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, NY.

STEIN, C. (1945). "A Two Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance". *Annals of Mathematical Statistics* 16, pp. 243–258.

WAN, H.; ANKENMAN, B. E.; and NELSON, B. L. (2005a). "Controlled Sequential Bifurcation: A New Factor-Screening Method for Discrete-Event Simulation". *Operations Research*, to appear.

WAN, H.; ANKENMAN, B. E.; and NELSON, B. L. (2005b). "Simulation Factor Screening With Controlled Sequential Bifurcation in the Presence of Interactions". Working Paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

WU, J. C. F. and HAMADA, M. (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*. Wiley, New York, NY.