LA-UR- 09-00030

Approved for public release;
distribution is unlimited.

Title: A Designed Screening Study with Prespecified Combinations of Factor Settings

Author(s): C.M. Anderson-Cook
T.J. Robinson

Intended for: Quality Engineering

A Designed Screening Study with Prespecified Combinations of Factor Settings

Christine M. Anderson-Cook
Los Alamos National Laboratory, Los Alamos, NM 87545
and
Timothy J. Robinson
University of Wyoming, Laramie, WY 82071

ABSTRACT

In many applications, the experimenter has limited options about what factor combinations can be chosen for a designed study. Consider a screening study for a production process involving five input factors that are extremely difficult to control. The goal of the study is to understand the effect of each factor on the response that is expensive to measure and destroys the part. From an inventory of available parts with known factor values, we wish to identify a best collection of factor combinations with which to estimate the factor effects. While the observational nature of the study cannot establish a causal relationship involving the response and the factors, the study can increase understanding of the underlying process. The study can also help determine where investment should be made to control input factors during production that will maximally influence the response. Since the factor combinations are observational, the chosen X-matrix will be non-orthogonal and will not allow independent estimation of factor effects. In this manuscript we borrow principles from design of experiments to suggest an 'optimal' selection of factor combinations. Specifically, we consider precision of model parameter estimates, the issue of replication, and ability to detect lack-of-fit. We present strategies for selecting a subset of factor combinations which simultaneously balance multiple objectives. The methodology is presented through a case study.

Keywords: Observational study, multiple objectives, D-optimality, screening, multicollinearity

INTRODUCTION

The ideal approach for exploring, understanding and establishing the causal connection between multiple factors and a response is through a designed experiment. However in many applications, factor levels cannot be manipulated by the experimenter and are thus observational in nature. Specifically, we consider a screening study for a production process involving five factors. An inventory of four hundred parts exists where each part is uniquely described by known level combinations of the five factors. The engineer would like to study the association of the response to the five factors but the response can only be measured after an expensive test requiring the destruction of the part. Given the cost involved with experimentation, only a small number of tests can be evaluated. Due to the sensitive proprietary nature of the application, values of the five factors have been altered but the original correlation structure has been preserved and is shown in Figure 1. Table 1 lists the correlation between the different factors. Clearly, moderate to large correlations exists among the factors and these correlations will have an adverse effect on the precision of estimation.

The available budget for the study allows for 20 parts to be selected and tested with the goal of estimating a first order linear model in the five factors. The measurement system had been previously evaluated and is known to not add much variability to the process. Since there is flexibility to select any subset of 20 parts from the available inventory of 400, the engineer desires to select an ideal set. The goodness of the selection is based on precise model estimation, ability to estimate pure error, and protection against model misspecification. The researcher wishes to rank the magnitude of the effects of the factors on the response. Since controlling the levels of the inputs is difficult, this information can be used to focus adjustments to the process to efficiently reduce in the variability of the response.

Sexton et al (2006) discuss a study which is part observational and part designed experiment where some factor levels are preset, but there are still options available for how different combinations should be created from the available data. While the set-up for our observational case study may appear to be somewhat specialized, there are other applications where optimal sampling from an existing population is desirable. For example, consider a financial application where there is interest in modeling the relationship between credit risk (the response) and various credit indices (the factors). Each potential customer is described by combinations of the credit indices but the response can only be measured via a detailed assessment of credit risk that is time-consuming and expensive. As with the production application, we would expect that that some of the financial measures (i.e. credit indices) are strongly correlated. In this case, the goal of a small study might be to select a small number of customers (i.e. distinct combinations of the credit indices) on which to do the detailed assessment in order to estimate the association between various financial measures and the desired credit risk response.

Similarly, consider an ecological application with interest in understanding the effect of different habitat markers on flora or fauna populations. Different locations with known environmental measures exist in a database and the ecologist can select a collection of locations to evaluate. Because of the labor- and cost-intensive nature of quantifying natural populations, only a small number of different locations can be considered. Some environmental measure combinations do not exist as actual habitat and many of the measures are highly correlated.

The general characteristics for problem under consideration are 1. difficulty in controlling input factor levels, 2. an existing collection of items with known input values from which to sample, and 3. an expensive, time-consuming and/or destructive measurement process which limits the number of samples which can be evaluated. In the next section we clarify the differences between the traditional design of experiments setting and the observational setting considered here.

DESIGNED TREATMENT SELECTION VS. SAMPLING OBSERVED TREATMENT COMBINATIONS

The optimal approach for collecting data for purposes of modeling is a designed experiment. For instance, a designed experiment to estimate a first order model for five factors of interest with 20 observations would perhaps suggest a 2^{5-1} fractional factorial with 4 center runs, as shown in Figure 2. This design allows for estimation of pure error, testing of lack of fit from both two-way interactions and quadratic terms, and has good D-efficiency. In what follows, we note some of the key features that distinguish this case study application from a standard screening designed experiment for 5 factors with 20 runs:

1. *Sampling involves an observational study versus experimenter control for setting factors in the designed experiment.* In our observational study with 20 data points selected from 400 candidate combinations of already specified input values, establishing causation is not possible. We can hope to understand the empirical relationship between the response and changes in the factor values but causation would need to be established with the active manipulation of the input values by the experimenter. This weaker association relationship is not ideal, but can still help provide guidance about what aspects of the inputs to focus on for future adjustments to the production process to reduce response variability.
2. *Correlation of factors cannot be eliminated when sampling observed treatment combinations.* As shown in Figure 1, many of the input factors are correlated, with some locations in a standard cuboidal design region not being possible. The correlations will mean that least squares estimates of the regression model parameters will not be independently estimated, as the columns of the X-matrix are not orthogonal. The geometry of many standard designs, such as the 2^{5-1} fractional factorial design, allow for independent estimation of factor effects. Given the correlation in the candidate set of design points shown in Figure 1, complete or even near orthogonality between the factors cannot be achieved. As a result, model parameter estimates will be dependent and less precise than from a designed experiment.
3. *Replicates not possible to estimate pure error when sampling.* Since the levels of the factors were not selected a priori and each part is characterized by a unique combination of the factors X_1-X_5, replication is not possible. In a designed experiment, the experimenter controls the settings of the factors and pure error variance is estimated from replicate experimental runs at one or more design points. An estimate of pure error variance is helpful to obtain a model free estimate of the natural variability of the underlying relationship between response and factors. For instance, if there are *m* locations with some

replication then

$$\hat{\sigma}^2 = MS_{PE} = {SS_{PE}}\Big/{df_{PE}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 \Big/ N-m,$$

where N denotes the total number of design points, and n_i represents the number of replicates at the i^{th} design point. SS_{PE} denotes the sum of squares for pure error, df_{PE} denotes the degrees of freedom for pure error, MS_{PE} denotes mean square pure error. In this application, there are no combinations of inputs with more than one observation. Since there are no pure replicates there is no direct way to estimate pure error.

4. *Lack-of-fit assessment in observational studies is problematic.* Related to the notion of replication is lack-of-fit. An important consideration when choosing a designed experiment is the ability to assess lack-of-fit. This provides information about whether additional terms in the model are suggested by the data. For example, suppose we assume a first order, but the true underlying relationship has one or more two-factor interactions or quadratic terms. Then the first order model will be inadequately describe the relationship and bias in estimation and prediction are likely. See Section 7.2.1 of Myers, Montgomery, Anderson-Cook (2009) for more details. To assess lack-of-fit, replicated combinations of factors are required that allow deviations from the predicted model form to be detected. Then the error sum of squares from the regression can be separated into estimates of lack-of-fit and pure error. The lack-of-fit can be quantified by comparing the model estimated response value with the average observed value at that location,

$$SS_{LOF} = \sum_{i=1}^{m} n_i\left(\bar{y}_i - \hat{y}_i\right)^2$$

where $\hat{y}_i$ denotes predicted response at the i^{th} design point from the fitted regression model. A popular approach for assessing lack-of-fit is then to use the F-statistic given by

$$F = \frac{SS_{LOF} \big/ (m-p)}{SS_{PE} \big/ (N-m)},$$

where N denotes the total number of experimental runs. For example, the 4 center runs in the design shown in Figure 2 allow for curvature from quadratic effects to be detected. However if no replication is possible, then $\bar{y}_i = y_i$ for each location and the error sum of squares cannot be divided into separate estimates of lack-of-fit and pure error.

5. *Location of the center of the candidate space is dependent on scaling choice in observational studies.* Replication of points at the design center are common and are referred to as *center runs*. When an experimenter defines the region of interest and the combinations of factors on which to collect data, there is an implicit assumption that by selecting the factor ranges, the experimenter believes that the scaled [-1,1] range roughly equates the different units of the factors in an equitable way. See Section 5.3 in Myers, Montgomery, Anderson-Cook, 2009 for more details. In observational studies where inherently there is no conscious selection of the ranges, this is not necessarily the case. In addition, when we have the choice of factor combinations to examine, the center runs are natural to locate at the scaled values of 0. With the lack of guaranteed symmetry of the factor values within the observed range, several logical options for the "center" of the candidate space might be considered. For example, we might select the mid-points of each factor range, $((X_{i,\min} - X_{i,\max})/2)$, to represent the center of the candidate space. Alternately, the set of factor means or medians may also be reasonable candidates for the "center". For experimental studies with uniform interest in all locations in the design space, these three potential choices would all be equivalent.

In this paper we discuss how to adapt various strategies employed in design of experiments to the sampling problem of selecting an ideal set of 20 factor combinations from the available 400. Specifically, we employ strategies which consider 1. precise estimation of model parameters; 2. estimation of pure error through pseudo-replicates and 3. assessing lack-of-fit for both two-way interaction and quadratic terms. It is important to note that in this case study, we are in a screening stages early in the process of understanding the relationship between response and factors, where the relative impact of the factors is important. Consequently, a first order model given by

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i x_i + \varepsilon \, , \qquad (1)$$

is appropriate, but it is possible that some curvature may exist. Hence, it is desirable to estimate the model parameters as precisely as possible, but the engineer would also like an estimate of pure error that is minimally model dependent. This will be helpful for testing the significance of the factor effects. Also of interest is to have a sample of points which allows for the ability to assess lack-of-fit.

SAMPLING STRATEGIES

For the experimental design setting, many software packages (SAS, Design Expert, MINITAB, SAS JMP, and others) can provide an optimal design based upon user-specified criteria and inputs. Typical inputs required from the user include 1. the design size (generally depends on the experimenter's budget); 2. the assumed underlying model form; 3. an objective function (such as D- or I-optimality) which relates to the user's goal for the data analysis; and 4. a candidate set of design points or a candidate design region from which the actual design points will be selected. For example, when screening of variables is of interest, popular objective functions are those which relate to the precision of the model parameter estimates. For more advanced stages of experimentation, the user is likely more interested in prediction and an objective function related to prediction variance will be specified. After the user has specified the four inputs described above, the software will provide the 'optimal' design for the specifications.

While the software may indicate that the design is 'optimal', it is important to remember that the good designs for any situation balance a wide array of attractive properties. Box and Draper (1975), and Myers, Montgomery, and Anderson-Cook (Section 7.1, 2009) suggest that several important qualities of a good design, including:

a. Results in precise estimates of model parameters.

b. Provides an estimate of "pure" error

c. Allows for assessment of lack-of-fit.

d. Allows models of increasing orders to be constructed sequentially.

e. Provides a check on the homogeneous variance assumption.

f. The design is cost effective.

We seek to identify a sample of 20 observations that perform well for the qualities above. Next, we describe several sampling strategies with varying emphases on the characteristics above.

Strategy 1: D-optimality for the first-order model

At the screening stage of experimentation, a popular strategy of determining an optimal experimental design is to choose a D-optimal design. The linear regression model can be written in matrix notation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the Nx1 vector of responses, $\mathbf{X}$ is the Nxp model matrix, $\boldsymbol{\beta}$ is the corresponding px1 vector of model parameters, and $\boldsymbol{\varepsilon}$ is the Nx1 vector of model errors, which are assumed to be i.i.d. $\mathrm{N}\left(0,\sigma^2\right)$. For our example, N=20 and p=6 upon considering the first order model in (1). The vector of least squares parameter estimates is then given by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ and the variance-covariance matrix of the vector of parameter estimates is

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2. \qquad (2)$$

The inverse of the variance-covariance matrix in (2), when scaled by the observation error variance, σ^2, and sample size, N, is known as the *scaled moment matrix*,

$$\mathbf{M}_s = \frac{\mathbf{X}'\mathbf{X}}{N}. \qquad (3)$$

D-optimality seeks to maximize the determinant of the scaled moment matrix. The resulting set of points will give the best precision of parameter estimates for the model specified.

When commercial software is used for optimal design selection, replicate factor combinations may be included for the optimal design. For the sampling problem of interest here, a single part corresponds to a unique candidate factor combination and thus replicates are not possible. Consequently, we have programmed a restricted exchange algorithm using the D-criterion for sample selection (outlined in the Appendix). Assuming the main effects model is correct, the D-optimal sample satisfies characteristic 'a' of a good design and can be as cost effective (characteristic 'f') as the sample size specified by the user. However, since the D-optimal approach does not consider the other characteristics, there are no guarantees on the performance of the sample for the other characteristics ('b'-'e'). Also, the results of this strategy may do poorly in terms of bias if the model is misspecified. Figure 3 shows the resulting D-optimal sample assuming a first order model.

Strategy 2: D-optimal sample for first order + two-factor interaction model

Similar to Strategy 1, we again utilize the D-criterion for treatment combination selection. Instead of assuming a first order model, we now assume the first order plus two-factor interaction model given by

$$y = \beta_o + \sum_{i=1}^{5} \beta_i x_i + \sum_{i=1}^{4} \sum_{i<j}^{5} \beta_{ij} x_i x_j + \varepsilon .$$

In this case, the model matrix **X** is a N x 16 matrix (1 intercept, 5 main effects, and 10 two-way interactions). Although the sample identified by this strategy will not produce the most precise model parameter estimates for the first order model, this strategy has a substantial advantage over Strategy 1. By considering the larger model, we are guaranteed that the sample will allow the user to estimate all main effects as well as two-factor interactions. Since Strategy 2 allows for estimation of the larger model, one can can assess lack-of-fit due to the presence of two-factor interactions. For Strategy 1, it is possible that the set of points produced by the selection algorithm will result in non-estimability of one or more of the two-factor interactions. Since we are in the screening stage of model selection, the existence of one or more two-factor interaction effects is possible. Figure 4 shows a pairs plot for the D-optimal sample for the first order + two-way interaction model.

While Strategy 2 offers an advantage over Strategy 1, neither strategy addresses the notions of replication and pure error. The ability to estimate pure error variance enables one to get a sense of the uncertainty in the experimental result when repeated tests are conducted on units exhibiting the same factor combination. Recall that true replication in this application is impossible since each part is uniquely associated with a specific factor combination. However, one can obtain a *pseudo* estimate of pure error variance if one is willing to assume that the underlying process model changes negligibly for small changes in the factor combinations. Specifically, one can consider the observed responses at distinct factor combinations $\mathbf{x}_i$ and $\mathbf{x}_j$, denoted by $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$, respectively, as *pseudo* replicates if one is willing to assume

$$E\left(y(\mathbf{x}_i)\right)-E\left(y(\mathbf{x}_j)\right)\approx 0 \text{ for } \left\|\mathbf{x}_i-\mathbf{x}_j\right\|\approx 0, \qquad (4)$$

where '$\| \;\|$' denotes the Euclidean distance.

Assuming (4) holds, we formulate two general strategies (Strategy 3 and 4) for sample selection. The first of these strategies considers *pseudo* replication at the center of the candidate factor combination space. In classical design of experiments, when factors are centered and scaled to have levels -1 to +1, the factor combination space is a multi-dimensional cube. Figure 2 provides a way of visualizing the space for five factors. With the 2^{5-1} design space, the locations of interest are assumed to be uniformly distributed throughout the cube. Uniformity implies that the center of the candidate region is the same whether one takes the midrange, the mean or the median of the factor levels. When the candidate space is observational, there is a non-uniform distribution of points throughout the candidate space. We consider three different methods for determining the center of the space: the midrange of the factor levels, the mean of the factor levels or the median of the factor levels. In each case the width of the scaled range is chosen to be 2 units, with the "center" being at a location of zero in each dimension. There is little to suggest a priori which method of centering will perform best based on the various criteria given the available 400 observations from which to select, so all three center measures were considered.

Strategy 3a: Five Center Runs using the Midrange + 15 Run D-optimal

Here, the minimum and maximum values of each X_i are set to -1 and +1, respectively. Therefore, the mid-point (X_1,X_2,X_3,X_4,X_5)=(0,0,0,0,0) is a natural choice for the center candidate space. Once the scaling is complete, we compute

$$\left\| \mathbf{x}_i - \mathbf{0} \right\| \tag{5}$$

for each of the i = 1,2,...,400 candidate points and choose the five factor combinations resulting in the shortest distance to **0**. After selecting these *pseudo center runs*, the exchange algorithm (described in the Appendix) is used to augment the center run set with fifteen additional factor combinations based upon the D-criterion assuming a first order model.

Strategy 3b and 3c: Five CR using the Mean and Median + 15 Run D-optimal

Here, the center of the candidate space (0,0,0,0,0) is taken to be either the means (Strategy 3b) or the medians (Strategy 3c) of each of the factors. Once the center has been determined, then the ranges of the scaled factors are adjusted to have width 2 units. This will likely produce a non-symmetric range around 0. As with Strategy 3a, we choose the five factor combinations resulting in the shortest distance to the center, and then used the exchange algorithm to augment the center runs with fifteen additional parts based upon the D-criterion and a first order model. Figure 5 shows the resulting sample from Strategy 3b.

The advantage of the different samples suggested by Strategies 3a, b and c is the ability to obtain a pseudo estimate of pure error, by considering several observations with similar locations near the center of the region of interest. If these points are sufficiently close in terms of Euclidean distance, then we can compute the sample variance among the five observed responses as a pseudo estimate of the natural process variability. In addition, these pseudo center runs can be used to quantify lack of fit due to second order model terms. Specifically, lack-of-fit can be quantified by comparing the average model estimated responses at these five points to the average observed values of the responses at these locations. The sum of squares lack-of-fit is then given by

$$SS_{LOF} = n_c\left(\bar{y}_c - \bar{\hat{y}}_c\right)^2 \tag{6}$$

where n_c denotes the number of center points and $\bar{y}_c$ and $\bar{\hat{y}}_c$, denote the average of the response values and model predicted values, respectively, corresponding to the sample center.

Strategy 4: Four Center runs + Four Pairs + 8 D-optimal runs

This strategy selects four observations closest to the center where the center is defined by any of the three measures described in Strategy 3. Next, four pairs of observations with minimal distance between points are selected. Note that the addition of four minimal distance pairs not only allows for improved estimation of the pure error but these pairs also allow some exploration of the assumption of homogeneous variance throughout the region of interest (characteristic 'e' of a good design). To select the four minimal distance pairs, we first identified the ten closest pairs and then randomly selected 4 pairs from these. The set of four center points and four minimal distance pairs were then supplemented with an additional 8 observations based on the D-criterion assuming a first order model using the exchange algorithm described in the Appendix. Multiple sets of 4 pairs of observations were selected as the starting point, and the sample with the best overall D-value was selected. Figure 6 shows the final selected sample using the midpoint of the factor ranges as the center.

SAMPLE COMPARISONS

We now present a comparison of the four described sampling strategies based on principles 'a' -'f' described earlier. Table 2 summarizes each strategy as it relates to the six characteristics of a good design. In the first two columns, the precision of model parameter estimates is addressed assuming a first order model. Specifically, relative D-efficiencies are provided for each strategy as well as the ranges on the standard errors, apart from σ, for the model terms. Note that the sample with the largest D-criterion of the four strategies is labeled as having a relative D-efficiency of 1 and all other relative efficiencies are computed relative to the D-value of the best sample. It is important to keep in mind that relative efficiencies are not calculated in terms of the ideal design in which factors are orthogonal to each other. The last row of the table labeled as "Ideal design" is for the set of factor combinations comprising the 2^{5-1} fractional factorial design with 4 center runs which might have been selected if controlling the input factor levels was possible.

While the first column can be helpful for a single number summary of each sample, the ranges for the standard errors for the model terms in the second column of Table 2 can be a more informative and more practical measure of performance. Figure 7a shows the values of standard errors for the models terms. The numerical values from the plot and the second column of Table 2 are multiplied by σ, the natural variability of the response. The first bar for each sample is for the estimated intercept while the remaining five bars are for the estimated coefficients for X_1-X_5. Note that for all samples, the precision of estimates is best for the intercept, and then for X_2 and X_4 terms (the third and fifth bars, respectively). These differences are dictated by the level of correlation between the factors for the selected samples. If we had been able to perform a designed experiment, then the standard error for each of the main effects would have been 0.25σ, where σ is the natural variability of the responses if input values are held fixed. Since we are not able to estimate σ until after the data are collected, we just focus on the relative size of the standard errors and ignore the constant multiplier σ from the comparison. Comparing 0.25 to what is possible with our samples, we see that all the sampling strategies have considerably larger standard errors for the factor main effects.

The sample from Strategy 1 (all 20 observations selected with the goal of optimizing D-efficiency for the first order model) is best in terms of D-efficiency. In terms of coefficient standard errors, the maximum standard error, 1.106σ, is only slightly better than Strategies 2 and 3. Strategy 4 has a substantially larger maximum standard error than the any of the other strategies. Note that there are some differences between the three samples obtained by considering different centers for Strategy 3, with the midrange design (3a) having the best relative D-efficiency (0.331) and the smallest maximum standard error, 1.2σ, for the individual factor effects.

In summary, if one is confident that the underlying model is a main effects model, Strategy 1 is best in terms of precision, but Strategies 2 and 3 perform comparably (characteristic ‘a’). Of the first three strategies, Strategy 3 is preferable since it is competitive based on precision of estimates while also offering the ability to estimate pure error (characteristic ‘b’), and an assessment of lack-of-fit (characteristic ‘c’). Of the three measures of center, the midrange is best in terms of maximum standard error of model coefficients.

Columns 3 - 5 of Table 2 compare the performance of the different samples assuming a larger underlying model with 16 terms (1 (intercept) + 5 (main effect) + 10 (two-way interactions)). Column 3 presents the relative D-efficiencies of each strategy, column 4 provides the ranges of the standard errors for the main effects and column 5 provides the ranges of the standard errors for the interaction terms. The results for the ideal design (2^{5-1} fractional factorial design with 4 center runs) are shown at the bottom of the table. The difference between the standard deviations for model term estimates for the ideal design and what is possible with our observational samples is even more dramatic with the larger model. Clearly the strong correlations between some factors are severely damaging our ability to estimate interactions. This makes intuitive sense, as to estimate an interaction term well, we must be able to see how the response changes when one factor is held constant and the other is changed. With high correlations between factors, this exploration is substantially hindered. Strategy 2 is the best in terms of D-efficiency and is used as the baseline for computing the relative D-efficiencies of the other strategies. Note that Strategy 2 dominates the other strategies in terms of precision The sample obtained using strategy 1 sample has only 3.7% relative D-efficiency and the other strategies have much smaller relative D-efficiencies.

Despite the vast differences in some of the relative D-efficiencies, it is helpful to compare based on standard errors of the main and two-way interaction terms. We consider the two groups of terms separately in columns 4 and 5 of Table 2. As with the D-efficiency summary, Strategy 2 performs best for the range of standard errors for the main effect terms with values between 0.87σ and 1.363σ. The standard errors associated with Strategy 1 are close in magnitude to those associated with Strategy 2 but Strategies 3 and 4 exhibit somewhat larger standard errors for the main effect terms.

Comparing columns 4 and 2 in Table 2, we note that the inclusion of the interaction terms in the model has increased the standard errors of the first order model terms. This is again the result of the correlations between factors from the non-orthogonal nature of the sample X-matrix. Note that for the ideal design, the orthogonal structure allows for the same standard error of 0.25σ to be preserved for all terms in both the first order model and the first order with interaction model.

When we consider estimating the two-way interactions, all strategies have large standard errors. This shows that the nature of the 400 observations from which we can choose severely limits our ability to estimate two-way interactions well. One advantage of this exploration of different samples is that it allows us to appropriately calibrate what is possible in the analysis phase before data are collected. In this case it should be clear that the lack of precision available to estimate the two-way interactions will preclude any formal testing of these terms unless the effects are extremely large or liberal p-values are considered for significance. Another option for exploring significance of these effects would be less formal graphical tools.

Although Strategies 1 and 2 produce samples that result in greater precision of the model parameter estimates (characteristic 'a'), its important to weigh this advantage against some of the more qualitative aspects of the samples. Strategies 3 and 4 place an emphasis on the ability to obtain an estimate of pure error (characteristic 'b'). The 5 pseudo center run samples of Strategy 3 have 4 degrees of freedom available for pure error estimation, while the 4 pairs and 4 pseudo center runs of Strategy 4 result in 7 (4 from pairs and 3 from pseudo center runs) degrees of freedom for this purpose. Strategy 4 with the 4 pairs distributed in different locations of the factor space also has some ability to check the assumption of homogeneity of variance throughout the region (characteristic 'e').

With the information summarized in Table 2, we are now in a position to evaluate which of the samples is best for our case study. The primary emphasis is on estimating the main effects for the 5 input factors, as this will help determine a strategy for where future resources should be spent to reduce the spread of one or more factors, which in turn might result in smaller variability in the response. The additional objectives of assessing two-way interactions or curvature from quadratic terms are precautionary in case the first order model is inadequate. The ability to do hypothesis testing is largely dependent upon on having a good estimate of the natural variability of the response, σ. Hence having an estimate that is less dependent on the assumed model and uses the pseudo-replicates is quite beneficial. Although Strategies 1 and 2 offer an advantage in the precision of parameter estimation, neither of these strategies allow estimation of pure error or an assessment of model lack-of-fit. While Strategy 4 offers the ability of estimating pure error, an assessment of lack-of-fit and the ability to check for homogeneity of variance, it does so at a severe cost for precise parameter estimation. Strategy 3 offers a nice balance of all properties (except for the ability to assess the homogeneity of variance assumption). The three measures of center are relatively similar for the first order model. Observing Figure 7b, the standard errors for the main effects in Strategy 3b appear to be somewhat smaller for the terms in the first order plus interaction model.

As a result of this comparison and discussion of the trade-offs, the engineer felt comfortable with the selection of the sample based on Strategy 3b, and was aware of what was realistically possible in the analysis phase.

CONCLUSIONS

In this paper, we have presented a case study of how to balance multiple objectives for a screening observational study where input factor values could not be controlled. The correlated structure of the data, and the inability to select and set input values reduces the precision of estimation of model terms, and leads to conclusions of associations between inputs and response, rather than the ideal conclusion of causality. The difference between the quality of estimation between the available samples and the ideal 2^{5-1} fractional factorial design with 4 center runs provides an important reminder of the benefits of planning and running a designed experiment whenever possible. In particular, the correlation structure makes a substantial difference if two-way interactions terms need to be included in the final model. As with this case study, there are times when a designed experiment is not possible and in these situations an observational study can still be helpful.

Planning a study should included selecting relevant characteristics from the lists provided by Box and Draper (1975) or Myers, Montgomery and Anderson-Cook (Section 7.1, 2009). Once these have been identified, it is possible to design selection strategies based on them, which will allow for careful comparison of the trade-offs between potential samples before a final sample is selected. In this particular study, focusing primarily on the first order model was thought to be sensible. However, the ability to check for inadequacies in this assumed model, both from two-way interactions or from quadratic terms, was also considered. In addition, the ability to estimate the natural variability of the response is informative as well as helpful for correctly calibrating any hypothesis tests for the terms in the model.

While planning the described strategies and writing the restricted exchange algorithms in the Appendix was a time-consuming process, the time and cost of developing these was still small in comparison to the total cost of collecting the data. By considering the relative performance of the different samples before a final one is selected, the engineer can gain an improved understanding of how well the characteristics of interest for the sample are satisfied and what estimation precision can be expected once the data are available. For the interested reader, the restricted exchange algorithms were in written in R (Venables et al, 2008) and are available from the authors upon request.

For observational studies, many of the well-known principles of good experimental designs can be adapted to help define the focus of different strategies for the sampling from available observations. Some elements, such as how to define the center of the "design space" need to be re-assessed. Some aspects of the analysis, such as estimating pure error and quantifying lack of fit, need to be redefined in this different setting where active manipulating factor levels is not possible.

Finally, although the engineer was restricted to only testing 20 observations from the available 400 parts, it can be helpful to explore samples that are slightly larger than the original plan. Table 3 shows the results of considering samples of size 25 for each of the four strategies. Although these samples are only 25% larger than the original size, the improvement in the precision of the parameter estimates, particularly when considering the two-way interaction terms, is substantial. This dramatic improvement may be in part due to the large number of terms in the first order with interaction model relative to the sample size of twenty. In a designed experiment setting, we would call this "nearly-saturated". Interestingly, the difference in relative efficiency and precision of parameter estimates between Strategies 1, 2 and 3 is less pronounced for the samples of size 25 compared to those of sample 20. This is due to additional observations being selected based on the D-criterion for each strategy. Based on the disproportionate improvement in the standard deviation of the model parameter estimates, the engineer was able to make a compelling quantitatively-based argument to request additional resources to expand the size of the study to size 25. Although in this case, expanding to the additional sample size was not possible because of funding restrictions, the exploration of a slightly larger sample size can still be a valuable exercise to understand the potential improvements and justify the value of the observations.

REFERENCES

1. Box, G.E.P. and Draper, N.R. (1975) "Robust designs", *Biometrika*, 62, 347-352.
2. Myers, Montgomery, Anderson-Cook (2009) *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: Wiley.
3. Sexton, C.J., Anthony, D.K., Lewis, S.M., Please, C.P., Keane, A.J. (2006) "Design of Experiment Algorithms for Assembled Products", *Journal of Quality Technology*, 38, 298-308.
4. Venables, W.N., Smith, D.M. and the R Development Core Team (2008) "An Introduction to R", www.r-project.org

APPENDIX

This appendix outlines the restricted exchange algorithms used for selecting the optimal sets of points for each of the sampling strategies discussed in the paper. Each algorithm assumes a candidate set of 400 factor combinations, C.

Sampling strategy 1: For the linear main effects model and a specified sample size, N, the algorithm proceeds as follows:

1. Randomly sample without replacement a sample of N treatment combinations, $\mathbb{S}$, from the candidate set.
2. Re-construct the candidate set to be $\{C_R = C - \mathbb{S}\}$.
3. Compute $D = |\mathbf{M}_s|$ where $\mathbf{M}_s$ (of equation (3)) is moment matrix based on a first order model for the sample $\mathbb{S}$.
4. Each candidate in C_R is considered as a replacement for each point in $\mathbb{S}$ in turn. By computing $D = |\mathbf{M}_s|$ for each potential candidate, the best replacement, if it exists, is found and inserted into $\mathbb{S}$ giving $\mathbb{S}^*$, the updated sample.
5. Re-construct the candidate set as $\{C_R = C - \mathbb{S}^*\}$ and repeat Step 4 with $\mathbb{S} = \mathbb{S}^*$ until there is no improvement possible in D.
6. Repeat Steps 1-5 for a large pre-defined number of random starts. The D-optimal sample has the largest value of D.

Sampling strategy 2: Same steps as outlined for Strategy 1 except in Step 3, $\mathbf{M}_s$ (of equation (3)) is moment matrix based on a first order plus two-factor interaction model provided.

Sampling strategy 3: For the first order model and a specified sample size, N, the algorithm proceeds as follows:

1. Scale the 400 observations based on an assumed center of the mid-point (or mean or median) of each factor.
2. Compute $\|\mathbf{x}_i - \mathbf{0}\|$ for each of the candidate factor combinations and choose the five factor combinations resulting in the smallest distance to **0**. Denote this set of points by $\mathbb{S}_{cr}$.
3. Re-construct the candidate set to be $\{C_R = C - \mathbb{S}_{cr}\}$.
4. Randomly sample without replacement a sample of N - N_{cr} treatment combinations, $\mathbb{S}_{\mathrm{D}}$, from the candidate set. The total sample of treatment combinations is now given by $\mathbb{S} = \mathbb{S}_{cr} + \mathbb{S}_{\mathrm{D}}$.
5. Re-construct the candidate set to be $\{C_R = C - \mathbb{S}\}$.
6. Compute $D = |\mathbf{M}_s|$ where $\mathbf{M}_s$ (of equation (3)) for the total sample $\mathbb{S}$.
7. Each point in $\mathbb{S}_{\mathrm{D}}$ is exchanged with each point in C_R in turn. For each exchange, the performance of the new sample is assessed by computing $D = |\mathbf{M}_s|$ where $\mathbf{M}_s$ is the moment matrix corresponding to the total sample $\mathbb{S}$. $\mathbb{S}_{\mathrm{D}}$ is then updated with the best exchange. Let $\mathbb{S}_{\mathrm{D}}^*$ be the updated sample.
8. Re-construct the candidate set as $\{C_R = C - \mathbb{S}_{cr} - \mathbb{S}_{\mathrm{D}}^*\}$ and repeat Step 7 with $\mathbb{S}_{\mathrm{D}} = \mathbb{S}_{\mathrm{D}}^*$ until there is no improvement in D.
9. Steps 3-8 are repeated for a large pre-defined number of random starts. The best sample from this strategy has the largest value of D.

Sampling strategy 4: For the linear main effects model and a specified sample size, N, the algorithm proceeds as follows:

1. Scale the 400 observations based on an assumed center for each factor.
2. Compute $\|\mathbf{x}_i - \mathbf{0}\|$ for each of the candidate factor combinations and choose the four factor combinations resulting in the smallest distance to **0**. Denote this set of points by $\mathbb{S}_{cr}$.
3. Re-construct the candidate set to be $\{C_R = C - \mathbb{S}_{cr}\}$.
4. Calculate the distance between all pairs of points in C_R, and identify the 10 pairs of points with the smallest distances, C_{Pairs}.
5. Randomly sample without replacement a sample of 4 pairs from C_{Pairs} to obtain $\mathbb{S}_{Pairs}$
6. Re-construct the candidate set to be $\{C_R$: $C_R = C - \mathbb{S}_{Pairs} - \mathbb{S}_{cr}\}$.
7. Randomly sample without replacement a sample of 8 factor combinations from C_R to obtain $\mathbb{S}_D$ from the candidate set. The total sample of treatment combinations is now given by $\mathbb{S} = \mathbb{S}_{cr} + \mathbb{S}_{Pairs} + \mathbb{S}_{\mathrm{D}}$.
8. Re-construct the candidate set to be $\{C_R$: $C_R = C - \mathbb{S}\}$.
9. Compute $D = |\mathbf{M}_s|$ where $\mathbf{M}_s$ (of equation (3) for the total sample.
10. Each point in $\mathbb{S}_{\mathrm{D}}$ is exchanged with each point in C_R in turn. For each exchange, the performance of the new sample is assessed by computing $D = |\mathbf{M}_s|$ where $\mathbf{M}_s$ is the moment matrix corresponding to the total sample $\mathbb{S}$. $\mathbb{S}_{\mathrm{D}}$ is then updated with the best exchange. Let $\mathbb{S}_{\mathrm{D}}^*$ be the updated sample.
11. Re-construct the candidate set as $\{C_R$: $C_R = C - \mathbb{S}_{cr} - \mathbb{S}_{Pairs} - \mathbb{S}_{\mathrm{D}}^*\}$ and repeat Step 10 with $\mathbb{S}_{\mathrm{D}} = \mathbb{S}_{\mathrm{D}}^*$ until there is no improvement in D.

12. Steps 5-11 are repeated for a large pre-defined number of random starts. The best sample from this strategy has the largest value of D.

Figure 1 Matrix of scatterplots showing pairwise candidate design points

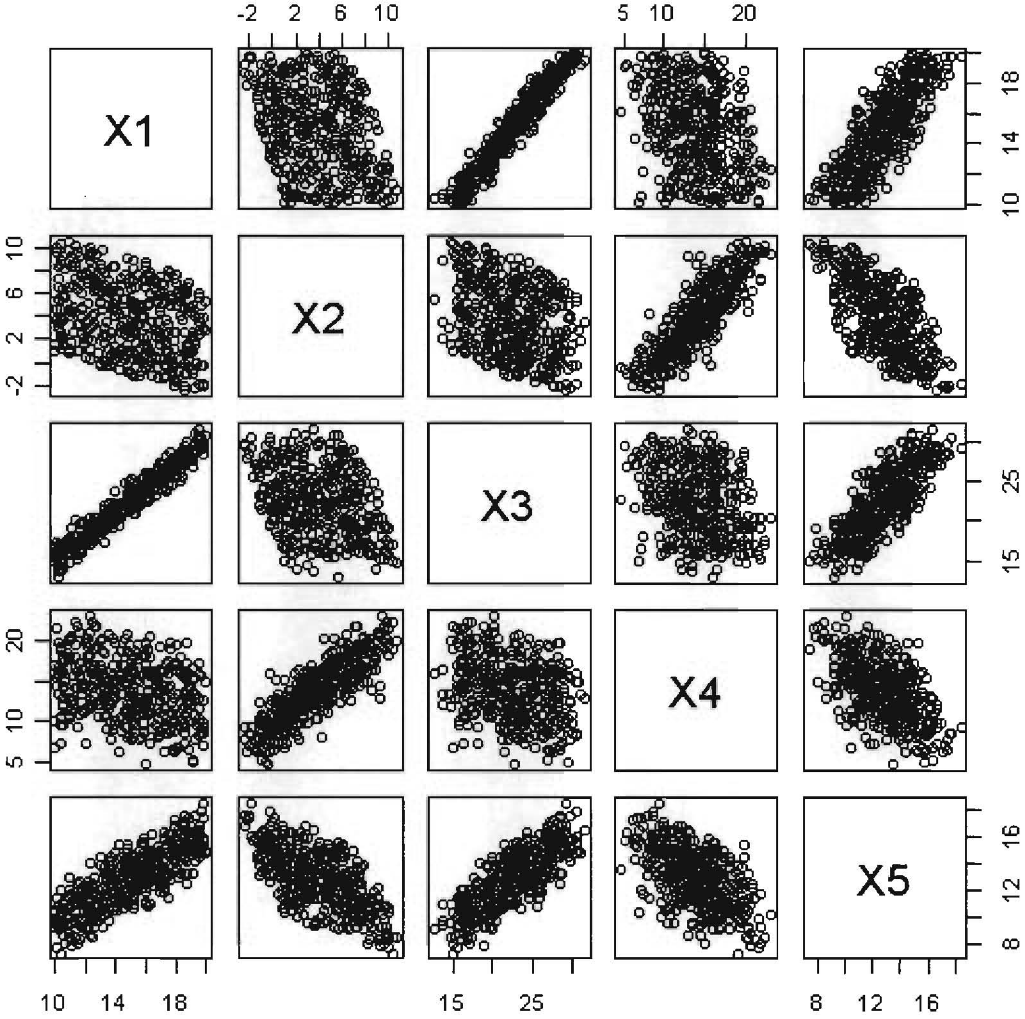

Figure 2: A standard 20 observation 2^{5-1}fractional factorial design

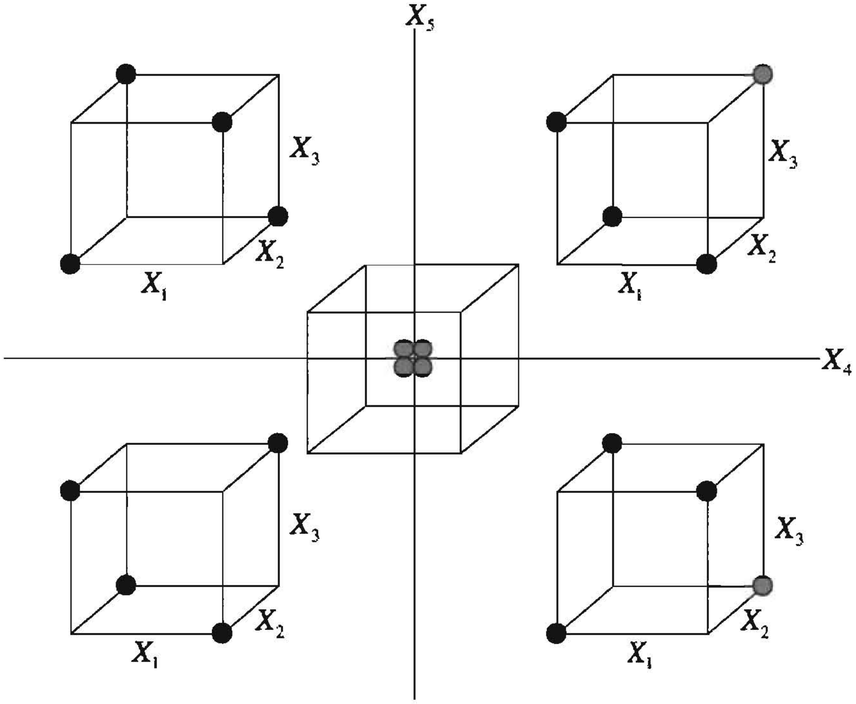

Figure 3: D-optimal sample assuming a first order model

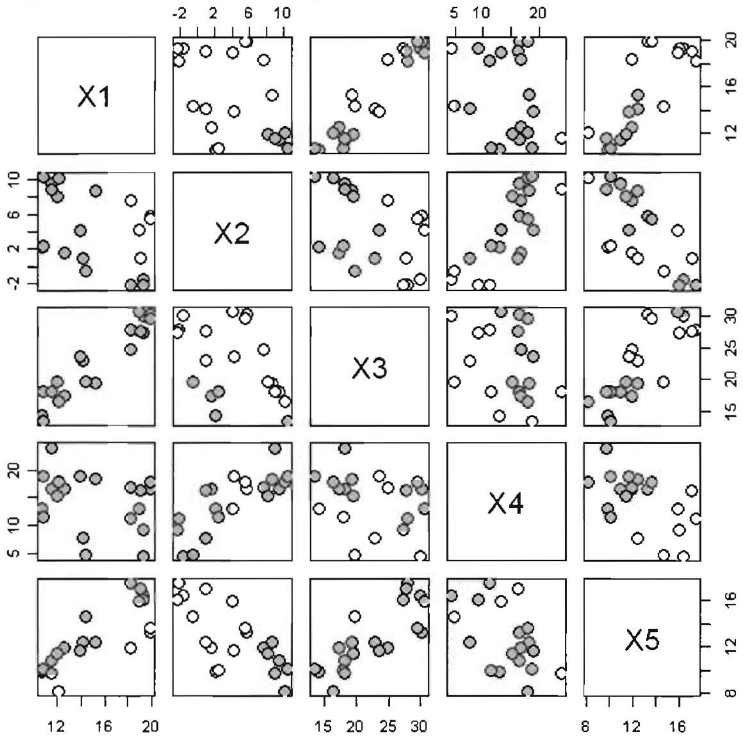

Figure 4: D-optimal sample assuming a first order plus two-way interaction model

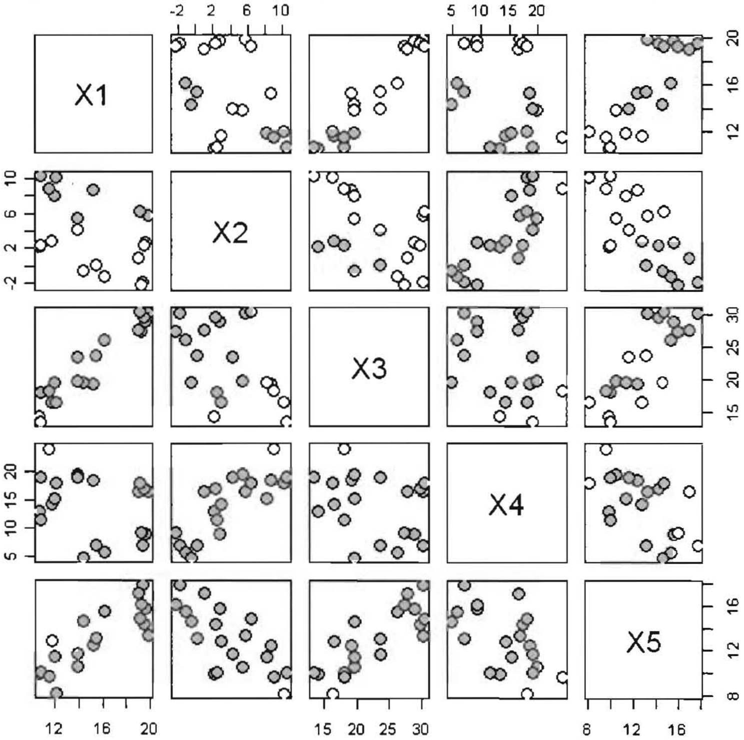

Figure 5: Best sample using Strategy 3b with the center based on the factor means with 5 pseudo center runs (black) and 15 observations based on D-optimality for a first order model (gray)

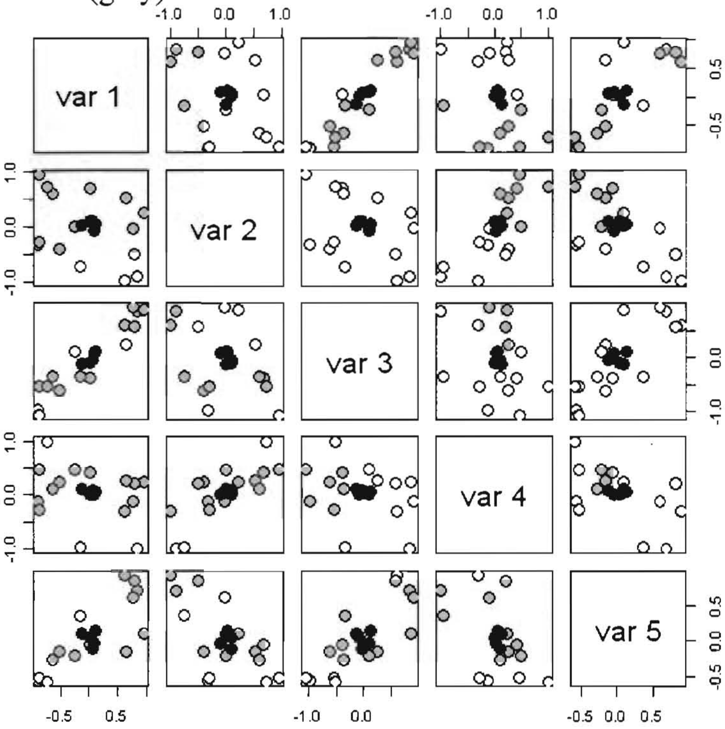

Figure 6: Best sample using Strategy 4 with 4 pseudo center runs (black), 4 pairs of points (white) and 12 observations based on D-optimality for a first order model (gray)

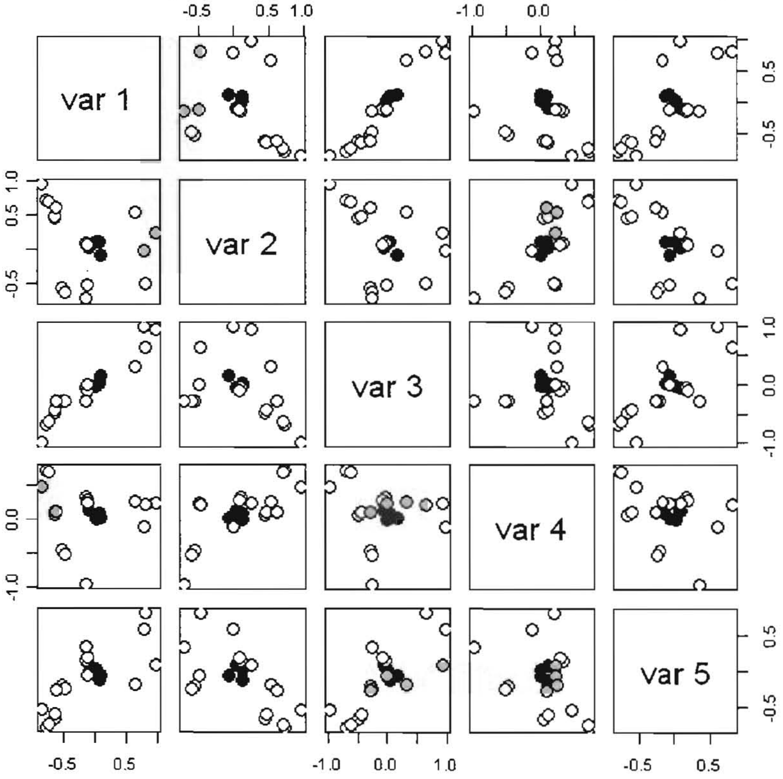

Figure 7: Barplot of standard deviations for 6 possible samples for different assumed models (a) Standard deviations for main effects based on assumed first order model, (b) Standard deviations for main effects based on assumed first order with interaction model

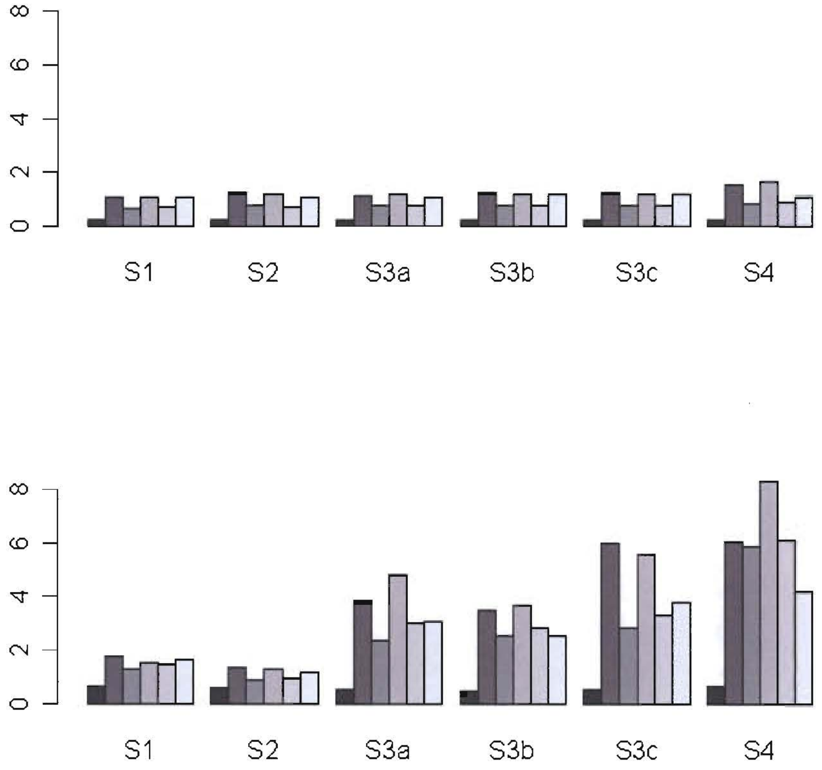

Table 1: Correlations between input factors for the 400 potential observations

	X1	**X2**	**X3**	**X4**
X2	-0.424			
X3	0.971	-0.445		
X4	-0.324	0.839	-0.347	
X5	0.837	-0.690	0.825	-0.577

Table 2: Comparison of different sampling strategies with 20 observations

		1st order		**1st + int**	**1st + int**		**Access**	**Access**
	1st order	**Beta_i**	**1st + int**	**Beta_i**	**Beta_ij**	**PE df**	**constant**	**quadratic**
Designs	**Rel D-eff**	**variance**	**Rel D-eff**	**variance**	**variance**	**(approx)**	**variance?**	**curvature?**
1. Dopt - first	**1.000**	0.684 - 1.106	0.04	1.280 – 1.742	1.835 – 11.413	0		
2. Dopt - 1 + int	0.620	0.738 - 1.257	**1.00**	0.870 – 1.363	1.295 – 5.302	0		
3a. 5CR - midrange	0.331	0.784 – 1.200	$\sim 10^{-5}$	2.362 – 4.783	2.811 – 33.966	4		Yes
3b. 5CR – mean	0.300	0.772 – 1.265	$\sim 10^{-5}$	2.505 – 3.653	2.541 – 29.581	4		Yes
3c. 5CR - median	0.296	0.772 – 1.267	$\sim 10^{-5}$	2.849 – 5.969	3.090 – 47.010	4		Yes
4. 4CR + 4pair + 8	0.100	0.845 – 1.663	$\sim 10^{-8}$	4.206 – 8.281	5.783 – 44.519	7	Yes	Yes
Ideal design		0.250		0.250	0.250	3		Yes

Table 3: Comparison of different sampling strategies if 25 observations had been possible

		1st order		**1st + int**	**1st + int**		**Access**	**Access**
	1st order	**Beta_i**	**1st + int**	**Beta_i**	**Beta_ij**	**PE df**	**constant**	**quadratic**
Designs	**Rel D-eff**	**variance**	**Rel D-eff**	**variance**	**variance**	**(approx)**	**variance?**	**curvature?**
1. Dopt – first	**1.000**	0.611 – 1.036	0.098	0.905 – 1.378	1.480 – 5.780	0		
2. Dopt - 1 + int	0.666	0.677 – 1.097	**1.00**	0.759 – 1.216	1.174 – 4.678	0		
3a. 5CR - midrange	0.431	0.703 – 1.062	$\sim 10^{-3}$	1.688 – 1.888	1.646 – 12.730	4		Yes
3b. 5CR - mean	0.405	0.676 – 1.099	$\sim 10^{-3}$	1.224 – 1.480	1.606 – 11.063	4		Yes
3c. 5CR - median	0.400	0.674 – 1.128	$\sim 10^{-3}$	1.119 – 1.995	1.390 – 11.628	4		Yes
4. 4CR + 4pair + 13	0.120	0.790 – 1.333	$\sim 10^{-6}$	1.479 – 3.290	2.305 – 12.975	7	Yes	Yes