Analysis of Definitive Screening Designs: Screening vs. Prediction

MARIA L. WEESE
Miami University, Oxford, OH 45056

PHILIP J. RAMSEY
University of New Hampshire, Durham, NH 03824

DOUGLAS C. MONTGOMERY
Arizona State University, Tempe, AZ 85281

Abstract

The use of definitive screening designs has been increasing since their introduction in 2011. These designs are used to screen factors as well as to make predictions. We assert that the choice of analysis method for these designs depends on the goal of the experiment, screening or prediction. In this work we present simulation results to address the explanatory (screening) use as well as the predictive use of definitive screening designs. To address the predictive ability of definitive screening designs, we use two five-factor definitive screening designs and simultaneously-run central composite designs case studies on which we will compare several common analysis methods. Overall we find that for screening purposes, the Dantzig selector using the BIC statistic is a good analysis choice, however, when the goal of analysis is prediction Forward selection using the BIC statistic produces models with a lower mean squared prediction error.

Keywords: explanatory modeling; predictive modeling; best-subsets; Dantzig selector; Forward selection; test data

1. Introduction

Definitive screening designs (DSDs), introduced by Jones and Nachtsheim (2011) for screening in the presence of second-order effects, have recently become popular in industry (Erler et al. (2012), Ramsey et al. (2015)). Practitioners have begun to use a single DSD in place of the traditional low-resolution screening and response surface design combination, to reduce the experimental run requirement and

cost. Examples of the use of DSDs for prediction are given in Renzi et al. (2014), Libbrecht et al. (2015), Erler et al. (2012), Ramsey et al. (2015) and Hercht et al. (2015).

Shmueli (2010) discusses modeling choices dependent on the goal of the analysis: predicting or explaining. She defines *explanatory modeling* as modeling to discover a causal relationship and *predictive modeling* as modeling where the singular goal is to predict the response. For example, when analyzing a designed experiment where the goal is to build a model for prediction the principle of heredity might be ignored in favor of smaller standard errors. But when the goal is interpretation, a lack of heredity can lead to a model that is hard to interpret. Consider the example from Montgomery et al. (2005) of a 2^3 design with three replicates to study the effects on the life of a cutting tool. Analysis reveals that factors B, C and A*C are important (p-value<0.05) and the main effect A is not important (p-value=0.8833). If the model is fit ignoring effect heredity (B, C and A*C), the MSE=28.817 and the estimated variance of prediction at the design corners is 4.8028. If the model is fit that obeys effect heredity, (A, B, C and A*C) then MSE=30.296 and the estimated variance of prediction at the design corner is 6.3121. Inclusion of the insignificant term preserves effect heredity to the detriment of the prediction interval width, but possibly better accuracy. Shmueli (2010) on pg. 296 notes the difference between the experimental goals of factorial designs and response surface methods (RSM). She notes that factorial designs are "focused on causal explanation in terms of finding the factors that affect the response" and response surface methods are "aimed at prediction". DSDs are used for both screening and prediction purposes.

This duality leads to interesting questions when deciding on the analysis strategy. There is a different approach to an analysis when the goal is to find the important driving factors *(explanatory analysis)* as opposed to prediction (*predictive analysis*). Jones and Nachtsheim (2011) suggest an analysis of a DSD following the procedure of Hamada and Wu (1992) enforcing strong effect heredity.

This procedure will produce nice explanatory models; those models might not be the best in terms of prediction.

In this paper we investigate the analysis of definitive screening designs in two ways. First, we investigate different analysis methods for using DSDs as intended, for screening in the presence of second-order effects. Then we compare different analysis methods when the goal of using a DSD is to make predictions. We evaluate predictive ability in two ways: using *in-sample* metrics and *out-of-sample* metrics (Shmueli 2010). The best way to evaluate a predictive model is to use "*out-of-sample*" metrics, such as mean squared prediction error ($MSPE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$), calculated on a test dataset of size *m*. In the field of design of experiments, where the primary goal is to gain the maximum amount of information from the fewest possible observations, creating a test dataset is not typically feasible and thus "*in-sample*" metrics are used to evaluate prediction, i.e. R^2 or the PRESS statistic (Allen (1971)). We compare the *in-sample* predictive power of each analysis method using the PRESS statistic. We evaluate *out-of-sample* prediction ability of each analysis method with MSPE, calculated on a test dataset. The "test set" was created by simultaneously-running a central composite design (CCD) with a DSD. We treat the DSD as a "training" dataset and use it to the construct models and treat the CCD as a test dataset to compare the predictive ability of each model. In this manner we directly compare the accuracy of the predicted values generated from the model fit on the DSD against the actual responses from the CCD. The advantage is that this comparison is made under real experimental conditions, not using simulation.

2. Definitive Screening Designs

DSDs introduced by Jones and Nachtsheim (2011) have the following appealing properties: (1) each of the *k* factors has three levels, (2) all main effects are uncorrelated with two-factor interactions, (3) no two-factor interactions are completely confounded, (4) quadratic effects are completely orthogonal to

main effects and no quadratic effects are completely confounded with any two-factor interactions, (5) for DSDs in six or more factors, a full second-order model is estimable in three or fewer factors, and (6) the number of required runs (*n*) is only one more than two times the number of factors. Xiao et al. (2012) improved upon the designs of Jones and Nachtsheim (2011) showing that DSDs can be constructed by stacking two *m* x *m* conference matrices and adding a center run. They also showed how conference matrices can be used to construct orthogonal DSDs for an odd number of factors. Jones and Nachtsheim (2013) constructed DSDs to allow for the inclusion of any number of categorical factors, but the analysis of those designs are not discussed in this work.

The main complication in analyzing DSDs arises from the possibility that both interactions and quadratic effects are active since those effects are correlated. Jones and Nachtsheim (2011) recommend analyzing these designs using Forward stepwise selection with the AICc statistic as the stopping criterion. If quadratic and interaction effects are found to be active, they recommend using Best-subsets regression to identify any model confounding. Note that Jones and Nachtsheim (2011) enforce strong effect heredity in their implementation of Forward stepwise regression and for their implementation of Best-subsets regression they fit all possible models with 10 or fewer first- and second-order terms.

2.1 Previous Studies

Dougherty et al. (2015) compared the performance of a nine-factor DSD with that of a nine-factor fractional Box-Behnken design (FBBD) with respect to effect heredity and effect sparsity for four different cases and two different noise levels. They analyzed the nine-factor DSD with Forward stepwise regression enforcing strong heredity as recommended by Jones and Nachtsheim (2011) and the FBBD with a factor-based Backward elimination recommended by Edwards and Mee (2011). They find that when strong heredity is present, the DSD performs best but attribute that to the fact that the analysis

method enforces strong heredity. They note that the DSD struggles to find active quadratic effects. Jones and Nachtsheim (2011) provide a small simulation study using the computer generated DSDs with k = {6, 8, 10, 12} and compare them in terms of power. They show that power increases as the number of runs, *n*, increases, the main effects have high power, the power for interactions is less than that of the main effects, and the quadratic effects have the lowest power. We expand upon the analysis of Doughtery et al. (2015) and Jones and Nachtsheim (2011) by using the Dantzig selector, Best-subsets regression, and Forward selection. We do not enforce any heredity restrictions into the analysis methods. In a recent paper, Jones and Nachtsheim (2017) suggest that all DSD designs should be modified in their construction by adding two "fake" factors, subsequently adding four more runs to the standard n=2k+1 DSD design size. However, we only consider the traditional DSD sizes. Errore et al. (2017) also perform a simulation study analyzing DSDs but do not address predictive ability in their simulations. They recommend that standard DSDs can be used to identify terms as long as the number of active terms is less than n/2.

3. Analysis Methods

We implement each of analysis methods described in this section using two model selection criterion, the corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC), as both are used in the design literature. As we are discussing two different analysis strategies, causal experimentation versus predictive experimentation, the best criteria for model selection might vary based on the analysis goal. In the rest of the paper we assume a standard linear model for a set of *p* effect columns constructed from the *k* factor columns in a DSD. The model takes the form

$$y = X\beta + \varepsilon \qquad (1)$$

where *y* and ε are n x 1 vectors, β is a (*p*+1) x 1 vector, and the elements of the error vector are independent with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

The Gauss-Dantzig selector, introduced by Candes and Tao (2007) and first used to analyze supersaturated designs by Phoa et al. (2009), has been shown to be useful for identifying active factors in supersaturated designs in several recent studies (Weese et al. (2015), Draguljić et al. (2014) and Marley and Woods (2010)). The Dantzig selector, a shrinkage method especially useful for when p>n, chooses terms that are consistent with the data and satisfy

$$min\left\|\hat{\beta}\right\|_1 subject\ to\ \left\|X^t\left(y - X\hat{\beta}\right)\right\|_\infty \leq \delta. \qquad (1)$$

Use of the Gauss-Dantzig selector (a two stage procedure) requires the specification of two tuning parameters γ and δ. γ is a threshold parameter to determine an active factor and δ is the shrinkage parameter. We use the automated procedure described by Phoa et al. (2009):

1. Compute $\delta_0 = max\left|x_i^t y\right|$.
2. Solve equation (1) for each value of δ over the range $0 \leq \delta \leq \delta_0$.
3. Identify the active effects as those whose coefficient estimates are larger γ than for each value of δ.
4. Using the effects identified in step (3), fit a linear model and obtain a value for the model section criterion (i.e. AICc or BIC).
5. Choose the model that has the best value of the chosen criterion.

We set γ=1.5 based on the true effect sizes and random error present in our simulation scenarios consistent with Marley and Woods (2010).

Weese et al. (2015) and Marley and Woods (2010) have shown success in terms of power by implementing the automatic selection procedure using the BIC statistic while Draguljić et al. (2014) used the AICc statistic. Although a DSD is not a supersaturated design when all effects are considered the full model matrix will be supersaturated and for this reason Krishnamoorthy et al. (2015) used the Dantzig

selector to analyze no-confounding designs and Errore et al. (2017) use the Dantzig selector to study DSDs but did not use the BIC statistic as the section criterion. Other regularization methods that have been considered in the analysis of supersaturated designs i.e. LASSO, Elastic Net etc. However, Draguljić et al. (2014) showed that the Gauss-Dantzig selector outperformed the LASSO. Given the evidence in favor of the Gauss-Dantzig selector we chose to include it in our study.

We compare the performance of the Gauss-Dantzig selector with Best-subsets regression and Forward stepwise selection using either the BIC and AICc as the selection statistic. Forward stepwise selection has the advantage that it is widely available for use in many commercial statistical packages. Best-subset regression was recommended for use if model confounding is present and will serve as our benchmark with which to compare the Gauss-Dantzig selector, but we recognize that Best-Subsets regression is computationally burdensome. In the simulations that follow, we restrict the maximum number of effects in a model to be nine and we score the top 56 models of each size (for computational considerations). All analysis methods were run in R.

4. *Explanatory* and *In-sample* Prediction Evaluation

To get a direct comparison of the three analysis methods (Gauss-Dantzig selector, Forward Selection and Best-subsets regression) for screening important factors we performed simulation studies with k={5, 6, 7, 8, 9, 10} using DSDs constructed in JMP 11.2 which employs the constructions of Xiao et al. (2012). We use each analysis method with AICc and BIC as the selection criteria for a total of six analysis methods. Four different experimental scenarios were used to assess the Power (proportion of active factors identified), Type I error (proportion of incorrect inactive effects identified as active), False Discovery Rate (proportion of effects identified as active that are actually inactive) and the average number of active factors identified. We subsequently calculated the measures under various simulation conditions, including models generated with varying degrees of effect heredity, strong or weak. We

define strong heredity such that a two factor interaction can only be declared active if both corresponding main effects are chosen to be active and weak heredity if an interaction can be declared active given one of the involved main effects are active. We only consider strong heredity for the quadratic effects meaning the corresponding main effect must be active for the quadratic effect to be chosen as active.

The parameter τ in Table 1 defines the magnitude of the effect and the parameter *a* in Table 1 defines the number of each effect types that are active. For example, in the first simulation scenario where $a = floor(n/3)$ for the main effects, using the DSD with n=13, k=5 the true model contains four randomly assigned active main effects with a magnitude of ±6. The effect signs are assigned randomly. The single active quadratic effect and interaction effects are chosen based on strong heredity from the randomly assigned active main effects and have a magnitude of ±3. The magnitudes of the main and interaction effects in the first simulation scenario mimic conclusions by Li et al. (2006) and their empirical study of active effects in unreplicated full factorial designs. They found that main effects are larger in magnitude than two-factor interactions and the majority of all active effects are main effects. Scenarios 2 and 4 assign all effects the same magnitude with the difference being scenario 2 ensures strong effect heredity and scenario 4 ensures weak effect heredity to assign the true active effects. Scenario 3 simulates a system with strong second-order effects that dominate the main effects. Note that we make no adjustments in the analysis for the differences in the true models. To analyze the simulations, we will treat our 144 separate simulations results as responses to a 4 x 6 x 6 factorial design as we have 4 Scenarios, 6 DSD sizes and 6 analysis methods.

In all scenarios the columns are chosen randomly and according to the stated heredity of the true model. We assume that changing an active factor will produce a change in the response and that changing an inactive factor results in no change in the response. Consequently, all inactive effects are

assigned a coefficient of 0. For each of 1000 iterations, a response is generated according to equation (1) where $\varepsilon \sim N(0,1)$.

Table 1: Simulation scenario protocol where *a* is the number of columns chosen to be active and τ is the active effect magnitude. Effect direction (+/-) was randomly assigned.

	Main Effects		Quadratic Effects		Interaction Effects		
Scenario	*a*	τ	*a*	τ	*a*	τ	Heredity
1	n/3	6	1	3	1	3	Strong
2	3	6	3	6	3	6	Strong
3	n/4	3	2	9	2	9	Strong
4	3	6	3	6	3	6	Weak

To compare the analysis methods in terms of *in-sample* predictive ability we tabulated the number of times each analysis method found the model with the lowest PRESS statistic where PRESS $= \sum_{i=1}^{n}(y_i - \hat{y}_{i-1})^2$. The PRESS statistic is computed by omitting the i^{th} observation from the fitted models. Note that when methods find the same model, each method is given credit for finding the lowest PRESS statistic. We provide results of the *out-of-sample* prediction evaluation of each analysis method the next section. The simulation protocol is as follows. For each of 1000 iterations:

1. Generate the response, *y*, according to specified scenarios in Table 1 and using equation (1).
2. Fit a model using each analysis method using the response generated from step (1).
3. Tabulate simulation statistics (overall power, Type I error, False Discovery Rate (FDR), model size, and power for each effect type).
4. Calculate the PRESS statistic for the selected model by each method.
5. Rank the models by the PRESS statistics, lowest to highest.
6. Tabulate the number of times each method found the model with the lowest PRESS statistic.

4.1 *Explanatory* Evaluation

Figure 1 and Table 2 display the power, Type I error rate and FDR averaged over all four simulation scenarios. Forward selection using the BIC statistic has the highest overall average power (0.711) but at the cost of the highest overall average Type I error rate (0.285). The Gauss-Dantzig (Dantzig) selector using the BIC has the second highest overall average power (0.705) with a much lower overall Type I error rate (0.092). The Dantzig selector using AICc has the lowest overall Type I error rate (0.059), but also the second lowest average power (0.574). Although we are pointing out differences between the analysis methods, it should be noted that the differences in average power ranges only about 9%. Forward selection using BIC had the highest overall False Discovery Rate (FDR) of 0.48. The Dantzig selector using AICc and Best Subsets using AICc had the lowest FDR of 0.200 and 0.199 respectively, less than half of that of Forward selection with BIC.

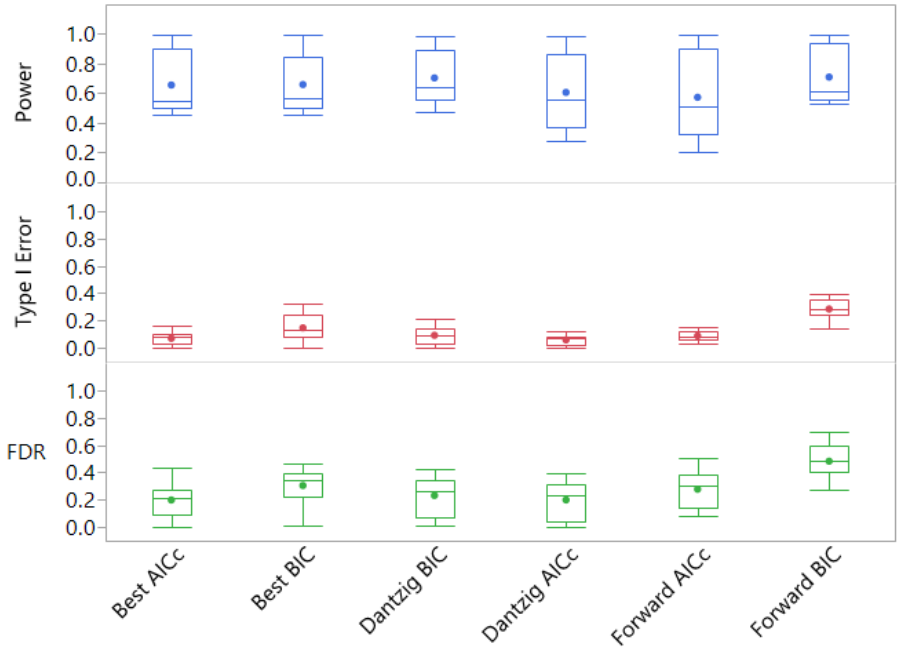


Figure 1: Overall power, Type I error and False Discover Rate (FDR) by method across all four scenarios

Table 2: Simulation results averaged over designs sizes and simulation scenarios.

Method	Average Power	Average Type I Error	Average FDR
Forward BIC	0.711	0.285	0.482
Dantzig BIC	0.705	0.092	0.232
Best BIC	0.660	0.147	0.304
Best AICc	0.657	0.07	0.199
Dantzig AICc	0.607	0.059	0.200
Forward AICc	0.574	0.088	0.278

Figure 2 displays the average power in each of the different simulation scenarios. Overall the DSDs perform best in the scenario 1 with limited second-order effects and dominant main effects. This is consistent with the findings of Jones and Nachtsheim (2011). There is little difference in power across all methods between scenario 2 (0.507), which generated the response using strong heredity in the true model, and scenario 4 (0.510) where the true model was generated using weak heredity. This is not surprising since our analysis methods did not take into account any effect heredity. Errore et al. (2017) did show an obvious gain in power when the true model is generated using strong heredity and the analysis method is forced to obey strong heredity. Scenario 3 contained dominant second order effects and fewer main effects than scenario 1 and had an average power of 0.633. The Dantzig selector using BIC had the highest average power in scenarios 2, 3 and 4 and the fourth highest power in scenario 1, although the average in scenario 1 was still 0.940.

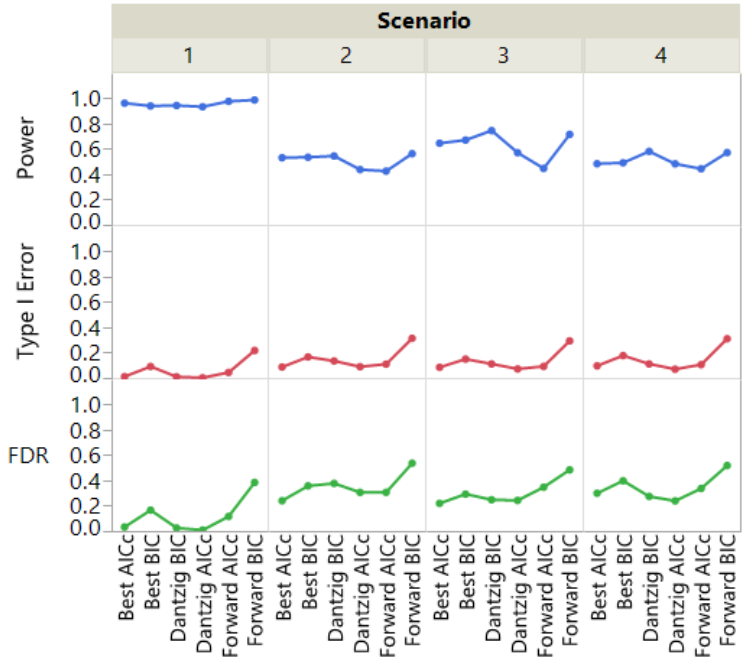
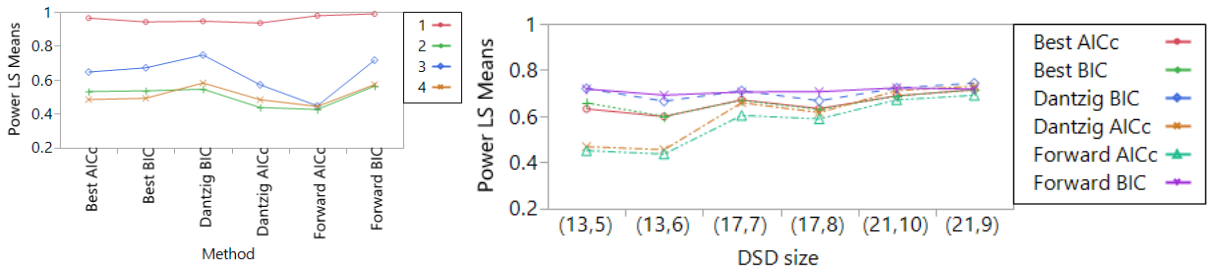


Figure 2: Power, Type I error and FDR by method by scenario. Note that the only difference between scenario 2 and 4 is strong vs. weak effect heredity in the generated "true" model.

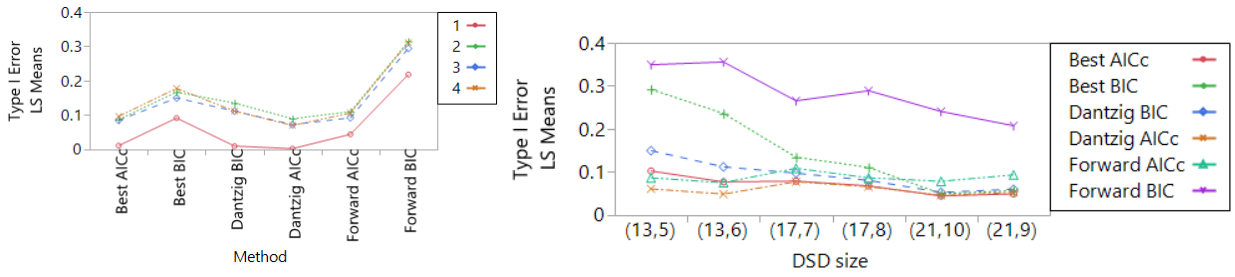
To strengthen our conclusions from the descriptive analysis we develop a model for power and treat Scenario, Method and DSD size (n,k) as factors fitting all main effects and two factor interactions. The model using raw power was ill-fitting via the standard residual analysis. As a result, we chose to model the square root of the number correctly identified active factors. This model (R^2=0.98) met the standard regression assumptions. The Method main effect and each interaction involving Method were found to be significant (p-value <0.0001). A Tukey multiple comparison test shows that Forward Selection with BIC and the Dantzig Selector using BIC have significantly higher power than the other methods but are not different from each other. Figure 3a shows an interaction plot of the Scenario*Method interaction. Interestingly, Forward selection with AICc gives virtually no difference in power between Scenarios 2, 3 and 4, where all the other methods show a similar pattern. The Dantzig selection using BIC performs slightly better in scenario 4 than 2. Figure 3b shows an interaction plot of

the Method by DSD size interaction and it is evident that the larger DSDs have more equivalent power by method than the smaller designs. It seems power does not become robust to analysis method until n>17. Similarly, a formal analysis of the square root of the Type I Error counts (R^2=0.99) with Method, Scenario and DSD size (n,k) as factors shows the Method main effect as well as the interactions involving Method to be significant (p-value <0.0001). A Tukey multiple comparison test shows that Forward Selection with BIC as the highest Type I Error rate and Best-subsets with AICc and Dantzig with AICc have the lowest Type I Error rate. Figures 4a and 4b show the interaction plots of the Scenario by Method and DSD size by Method interaction for the Type I error rate. Figure 4b shows that Forward selection with BIC has a higher Type I error rate of all the methods for even the larger design sizes. Lastly a formal analysis of the square root of the counts of False Discoveries (R^2=0.97) reveals all main effects and interactions involving Method as significant (p-value <0.0001). Figures 5a and b show the Method by Scenario and the Method by DSD size interaction plots. Most notably Figure 5b shows that for all methods, the FDR stays relatively flat with increasing design size.



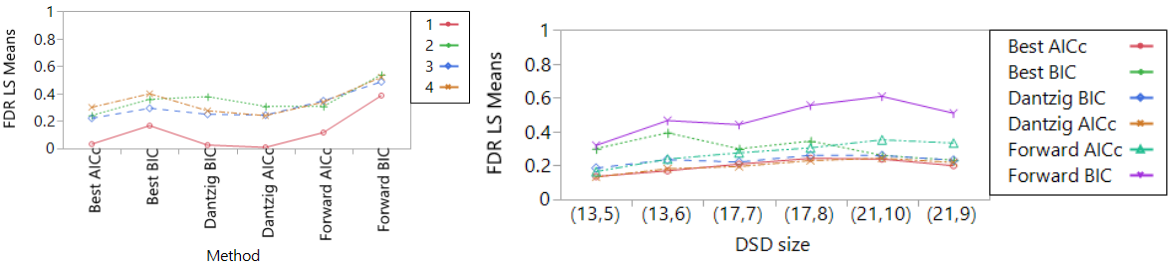
(a) **(b)**

Figure 3a and b: Interaction plots of the Scenario by Method Interaction (3a) and Method by DSD size interaction (3b) for a model with square root of Power counts as the response. Note the y-axis is adjusted to show Power.



(a) **(b)**

Figure 4a and b: Interaction plots of the Scenario by Method Interaction (4a) and Method by DSD size interaction (4b) for a model with the square root of the Type I error as the response. Note the y-axis is adjusted to show Type I Error Rate.



(a) **(b)**

Figure 5a and b: Interaction plots of the Scenario by Method Interaction (5a) and Method by DSD size interaction (5b) for a model with False Discovery Counts as the response. Note the y-axis is adjusted to show False Discovery Rate.

Figure 6 shows the breakdown of the average power by effect type (main effect, interaction, quadratic), scenario and analysis method. As expected the main effects have the highest average power, although it is notable that Forward selection and the Dantzig selector using AICc have the lowest average main effect power in scenarios 2, 3 and 4. The Dantzig selector using BIC has the highest average power to detect the correct interactions in scenarios 2, 3 and 4, see the top row of Figure 2. Forward selection with BIC seems to have the highest power to detect the quadratic effects. It is

notable in Figure 8 that Best-subsets regression using the BIC statistic tended to produce models with nine terms, which was set as constraint in the implementation of the method. To strengthen our conclusions as we fit a model using Method, Scenario and DSD size (n,k) for the three responses: the square root of Power count of Quadratic effects (R^2=0.98), the square root of Power count of Interactions (R^2=0.99), and the square root of the Power Count of Main Effects (R^2=0.97). The Method main effects and two factor interactions involving Method were highly significant (p-value <0.001) for the main and quadratic effects and significant (p-value<0.0215) for the interactions. A Tukey's multiple comparison test shows that the Dantzig selector and Forward Selection using AICc have significantly lower power to detect the main effects, the Dantzig selector using the BIC has significantly higher power to detect the Interaction effects and that Forward Selection with BIC has the highest power to detect the quadratic effects and Dantzig using the AICc has the lowest.

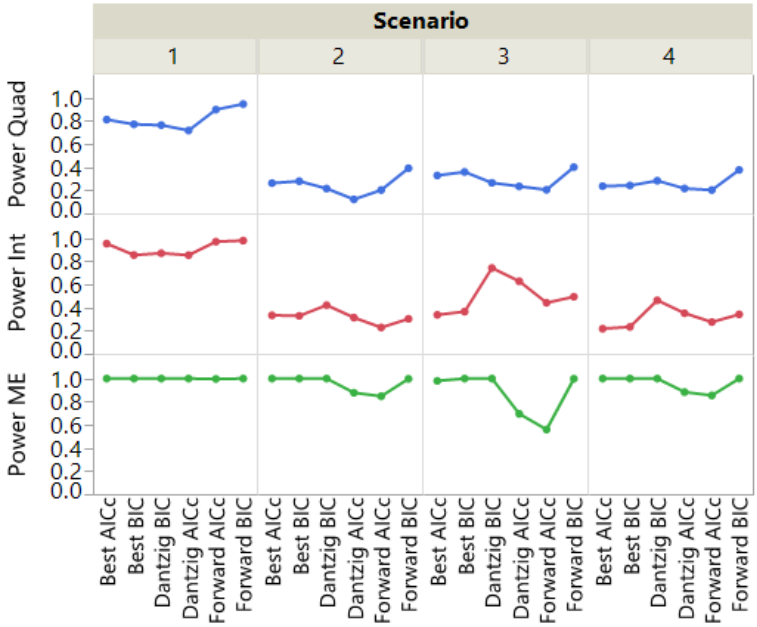
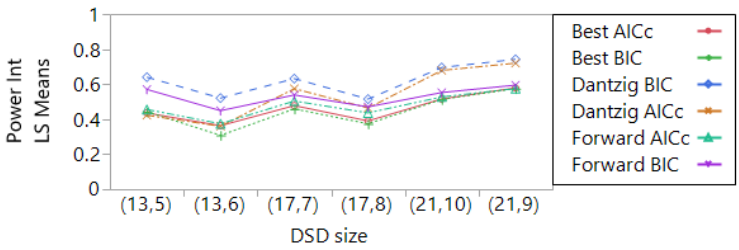
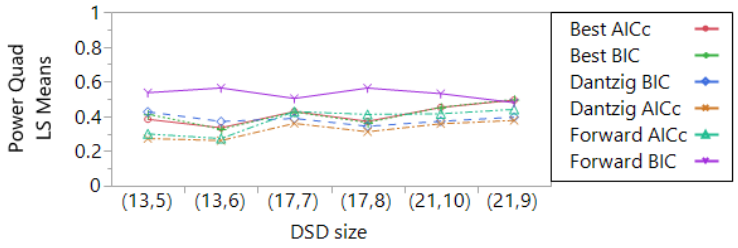


Figure 6: Power based on effect type across all four scenarios.

The interaction plot between Method and DSD size when the power of the main effects is the response looks very similar to Figure 3b indicating the larger designs sizes are fairly robust to the choice of analysis method. However, the interaction plots of the Method by DSD size interaction when the response is the power to detect the interactions (Figure 7a) and the power to detect the quadratic effect (Figure 7b) show a less dramatic difference in the power as the design size increases.



(a)



(b)

Figures 7a and 7b: Interaction plots of the Method by DSD size for y= square root of the power counts for interactions (7a) and Method by DSD size interaction (7b) for a model with y=square root of the power counts for the quadratic effects. Note the y-axes have been adjusted to show Power.

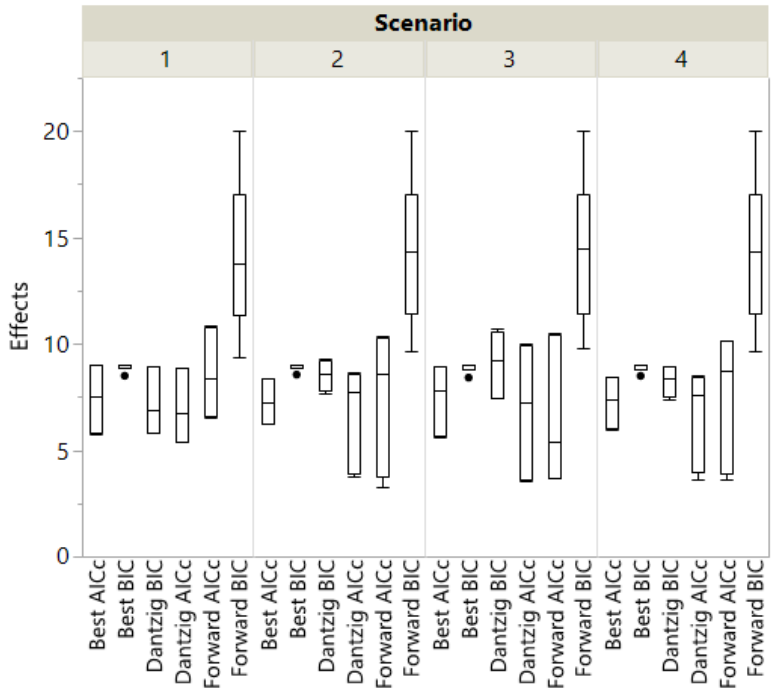
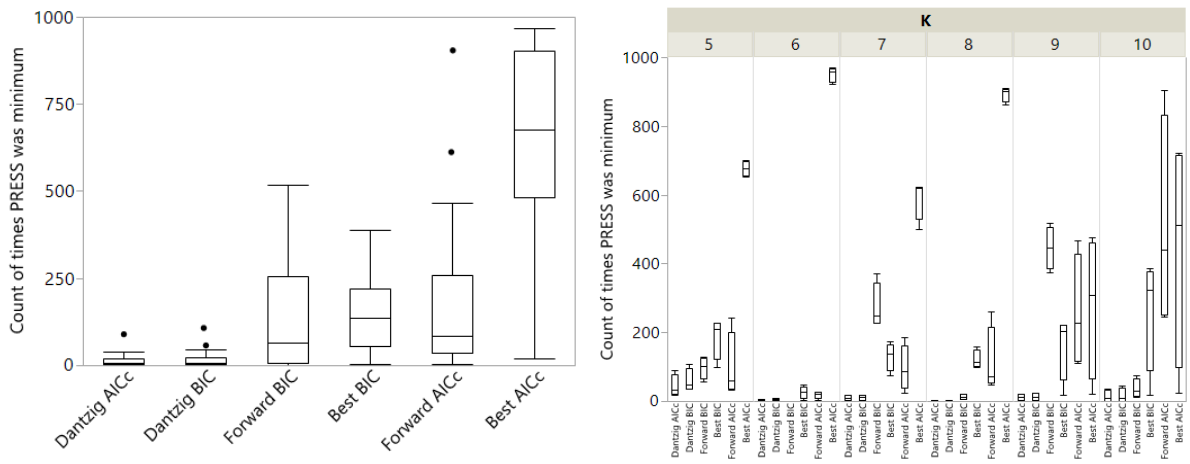


Figure 8: Average number of effects found by method for each scenario.

4.2 *In-Sample* Prediction Evaluation

It is interesting that the methods with the highest power were not also the methods that had the lowest PRESS. Best-subsets regression using AICc was the analysis method that most often had the lowest PRESS statistic (see Figure 9a) indicating that the goal of the experiment might be considered when choosing an analysis method. Best-subsets using AICc was clearly dominant for the smaller designs, but as *k* increased (and subsequently *n*) the effect dissipated (Figure 9b).



(a) **(b)**

Figures 9a and b: The number of times each method had the lowest press score across design size and scenario. And the minimum PRESS over scenario by k. Note, ties are included.

5. *Out-of-Sample* Prediction Evaluation

To evaluate prediction using *out-of-sample* methods we use a test dataset. In the field of designed experiments, test data is typically not available. It is standard practice to perform confirmatory runs to verify process settings or a process optimum, but it is not standard to use those runs to select the model. Model selection is typically done using *in-sample* metrics. In the following two case studies we have access to test datasets which are simultaneously-run central composite designs, using the same factors as each DSD and run under similar experimental conditions. In a typical predictive modeling procedure, the model is selected based on the performance on test data. In this case we are not using test data to aid in model construction but as a means to evaluate the models constructed by the different analysis methods using the DSD. We will use the CCDs as a test dataset and evaluate each of the models chosen on the DSD by the different analysis methods based on mean squared error of prediction (MSPE). Note the predicted values in the MSPE are generated from the design points of the CCD using the models found by analyzing the DSD. We determine the best model as the one which has

the lowest $MSPE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ on the test dataset (in this case the CCD) where *m* is the number of samples in the CCD.

The goal of the first experiment (Ramsey and Ornek (2013)) is to optimize the biomolecule yield of the fermentation step in a bio-process. The k=5 factors used are pH of the fermentation solution (pH), Dissolved Oxygen (%DO), Induction Temperature, Induction OD600 (biomass at which the induction is initiated as measure by optical density at 600 nm), and Feed Rate of a growth media. This experiment was performed prior to the construction recommendations of Xiao et al. (2012) and as such the design used was constructed according to Jones and Nachtsheim (2011) with n=11 unique treatment combinations. Additionally, four center point replications were added to obtain an error estimate. The corresponding CCD had n=26 experimental runs with five additional center points for a total of n=31 runs with axial runs of ±1.3 chosen based on process knowledge. The second case study is also a k=5 factor DSD, but using Xiao et al. (2012). The goal of the experiment is to optimize an analytical method for glycoprofiling. Glyocoprotiens are the largest group of biologically-derived drugs and currently there is a lack of a universal analysis technique for glycosylation analysis. The five factors in this experiment are Initial %NaOAc (%A), Initial %NaOH (%B), Gradient_01 (mM NaOAc/min), Gradient_02 (mM NaOAc/min) and Gradient_03 (mM NaOAc/min). The corresponding face centered CCD had 28 experimental runs and face centered axial runs. As typical in a predictive modeling situation, we fit models using the training dataset, the DSD, and compare the models using the test dataset, the CCD. The six different models will be generated by the different analysis methods and selection criteria. The procedure is a follows:

1. Using one of the analysis method/selection statistic combinations (Dantzig selector, Forward Selection and Best subsets regression), find the best model using the responses and factor combinations of the DSD.

2. For each of the six models, generate $\hat{y}_i$ values using the design points of the CCD. Note for the Dantzig selector we will fit a standard linear model to obtain the coefficient estimates.
3. Calculate the MSPE as in equation (3) using the $\hat{y}_i$ values from step (2) and the true responses from the CCD as y_i.
4. Rank the analysis methods according to lowest MSPE.

Table 3: Summary of the *out-of-sample* prediction evaluation for each analysis method by average squared error for the Fermentation Experiment.

Method	Model	MSPE on CCD
Forward Selection BIC	X5 X1:X2 X2 $X3^2$ X1 X3:X5 X4 X3 X3:X4	2997.956
Best-subsets Regression BIC	X1 X2 X3 X4 X5 $X5^2$ X2:X5 X3:X5	3930.691
Best-subsets Regression AICc	X1 X2 X5 $X5^2$ X3:X5	4464.483
Dantzig BIC, γ=1.5	X1 X2 X3 X4 X5 $X1^2$ $X2^2$ $X3^2$ $X4^2$ $X5^2$	4688.713
Forward Selection AICc	X5 X1:X2 X2	6924.568
Dantzig AICc, γ=1.5	*null*	13721.883

Table 4: Summary of the *out-of-sample* prediction evaluation for each analysis method by average squared error for the Glyoprofiling experiment.

Method	Model	MSPE on CCD
Forward Selection BIC	X1 $X1^2$ X1:X2 X2 X4 X1:X3 X5 X3:X4	0.06376
Forward Selection AICc	X1 $X1^2$ X1:X2 X2 X4 X1:X3	0.06449
Best-subsets Regression BIC	X1 X2 X4 X5 $X1^2$ X1:X2 X1:X3	0.06726
Dantzig, BIC γ =0.5	X1 X2 $X1^2$ $X2^2$ $X3^2$ $X5^2$ X1:X2 X2:X3	0.07328
Best-subsets Regression AICc	X1 X2 X4 $X1^2$ X1:X2	0.08057
Dantzig AICc γ=0.5	X1 $X1^2$ $X2^2$ $X3^2$	1.2445
Dantzig, BIC γ=1.5	X1 $X1^2$ $X5^2$	1.2448
Dantzig AICc γ=1.5	X1 $X1^2$ $X5^2$	1.2448

The method that seemed to have the best *in-sample* prediction evaluation by the PRESS statistic, Best-subset regression using AICc, does not have the best *out-of-sample* prediction evaluation. In both case studies Forward selection using BIC gives the lowest value of the MSPE on the test dataset. Forward selection using BIC did perform well in the power simulations, but it had a large Type I error

rate. Although a high Type I error rate might be acceptable in a screening situation a method with similar power and a lower Type I error rate is preferable. The best screening or *explanatory* method, Dantzig selector using the BIC statistic, did not perform as well in the *out-of-sample* prediction evaluation, see Tables 3 and 4. Note that for the Gauss-Dantzig selector different values of γ will change the model selected by the Gauss-Dantzig selector. A smaller value will allow more terms to be considered at each value of δ. For the Fermentation experiment, lowering the value of γ does not change the number of terms considered at each value of δ since all coefficient estimates are larger than 1.5. For the Glycoprofiling experiment a smaller value of γ is preferred due to the magnitude of the coefficient estimates. For further guidance on choosing γ see Candes and Tao (2007). Both experiments show disagreement amongst the chosen models.

Thus far we have compared the average prediction performance of each model selection method using in-sample or out-of-sample methods. Practitioners are often concerned with the ability to predict a single point in the design region (i.e. the predicted minimum or maximum) and as such we evaluated the DSD and model selection method on the individual prediction at the maximum response using the Gycloprofiling and Fermentation experiments. We use the maximum predicted value from the full second order model fitted to the CCD as the metric against which to judge the quality of prediction from the DSD and model selection method. Tables 5 and 6 give the individual values and the 95% prediction interval for the DSD model selection method and the benchmark, the predicted value from the CCD. Interestingly, there is little disagreement between MPSE calculated using the design points from the CCD and the models found on the DSD (Tables 3 and 4) and the individual predictions at the maximum (Tables 5 and 6). The only exceptions (noted by the changed ordering in Table 6 as compared to Table 4) are for Best subsets Regression using BIC and the Gauss-Dantzig selector using AICc with γ =0.5. Perhaps this is due to the fact that in this experiment several large, dominant effects were

present. Overall, the methods that produced the best prediction on average, also produced the best individual prediction for these two experiments.

Table 5: Predicted maximum response value using the reduced models on the DSD compared to the full second order model fit to the CCD for the Fermentation experiment.

Method	Model	Predicted Max Response	95% PI of Max Predicted Response
Using CCD	Full Second order model	663.727	512.46, 814.994
Forward Selection BIC	X1 X2 X3 X4 X5 $X3^3$ X1:X2 X3:X5 X3:X4	606.924	526.979, 686.869
Best-subsets Regression BIC	X1 X2 X3 X4 X5 $X5^2$ X2:X5 X3:X5	522.681	462.878, 582.484
Best-subsets Regression AICc	X1 X2 X5 $X5^2$ X3:X5	509.7855	454.358, 565.213
Dantzig BIC, γ=1.5	X1 X2 X3 X4 X5 $X1^2$ $X2^2$ $X3^2$ $X4^2$ $X5^2$	483.6595	425.117, 542.202
Forward Selection AICc	X2 X5 X1:X2	519.0221	437.198, 600.846
Dantzig AICc, γ=1.5	*null*	-	-

Table 6: Predicted maximum response value using the reduced models on the DSD compared to the full second order model fit to the CCD for the Glycoprofiling experiment.

Method	Model	Predicted Max Response	95% PI of Max Predicted Response
Using CCD	Full Second Order Model	13.88527	13.0304, 14.7401
Forward Selection BIC	X1 X2 X5 X4 $X1^2$ X1:X2 X1:X3 X3:X4	13.82082	13.1951, 14.4465
Best-subsets Regression BIC	X1 X2 X4 X5 $X1^2$ X1:X2 X1:X3	13.75837	13.1612, 14.355
Forward Selection AICc	X1 X2 X4 $X1^2$ X1:X2 X1:X3	13.58997	12.9681, 14.2118
Dantzig, BIC γ =0.5	X1 X2 $X1^2$ $X2^2$ $X3^2$ $X5^2$ X1:X2 X2:X3	13.3949	12.2477, 14.3132
Best-subsets Regression AICc	X1 X2 X4 $X1^2$ X1:X2	13.33993	12.7743, 13.9055
Dantzig, BIC γ =1.5	X1 $X1^2$ $X5^2$	11.94555	10.3429, 13.5482
Dantzig AICc γ =1.5	X1 $X1^2$ $X5^2$	11.94555	10.3429, 13.5482
Dantzig AICc γ =0.5	X1 $X1^2$ $X2^2$ $X3^2$	11.46669	9.53725, 13.3961

6. Discussion and Conclusions

We have shown that the choice of analysis to use for an experiment run with a DSD is dependent on the application of the DSD as either a screening design or a design for prediction. We should note that successful use of a DSD in place of a larger response surface and screening design combination requires that practitioners have defined the correct experimental region. Foregoing the initial steps of screening and steepest accent does increase the risk of not finding the optimum response. In many ways, the dual usage of a DSD makes it a great design to consider for an initial experiment, although we have also shown the conclusions from a smaller DSD will be more dependent on the choice of analysis method than those from a larger DSD. In terms of *explanatory* modeling or screening, we can recommend the Dantzig selector using the BIC statistic for several reasons. First, it had some of the highest overall power for the different effects types as well as lower Type I error rates. The Dantzig selector using BIC is going to have better power to detect the interaction effects when the true model is dominated by second-order effects or the magnitude of those effects is similar to the main effects. When the application of a DSD will involve prediction, we recommend analyzing the results with Forward selection using the BIC statistic. Forward selection using the BIC statistic also is nota poor choice for screening, but one can expect high Type I error rates. Overall, the methods with the BIC statistic seemed to perform better with the exception of the *in-sample* prediction evaluation in which case Best-subsets regression using the AICc statistic was found to have the lowest PRESS in the majority of cases considered. We believe that the differences between the analysis methods are due to the difference in experimental goals. However, as suggested by a reviewer the differences between the analysis methods could be due to the fact that the true models for the Fermentation and Glycoprofiling experiments are not as sparse as the models we presented in our simulations. When one suspects the true model will not have sparsity, it might be prudent to consider the addition of "fake factors" (Jones,

B. and Nachtsheim, C.J. (2017)) or design augmentation. Additionally, it should be noted that in the simulations we summarized the prediction performance on the individual design points.

The DSDs seem to perform best when the true model is dominated by strong main effects and has relatively few second-order effects. We did not investigate how the power could be improved by adding heredity restrictions into the model selection method. It is fairly obvious that since the main effects have the highest power, if the true model has strong heredity present, introducing that restriction would increase the power to detect the second-order effects. Future work on DSDs could include augmentation if they are used as a screening experiment and consideration of multiple solutions for prediction evaluation.

References

Allen, D.M. (1971), "Mean square error of prediction as a criterion for selecting variables", *Technometrics*, 13, 469-475.

Candes, E. and Tao, T. (2007) "The Dantzig selector: statistical estimations when p is much larger than n". *The Annals of Statistics*, 35(6), 2313-2351.

Draguljić, D., Woods, D.C., Dean, A.M., Lewis, S.M., and Vine, A. J. E. (2014) "Screening strategies in the presence of interactions". *Technometrics*, 56(1), 1-16.

Dougherty, S., Simpson, J.R., Hill, R.R., Pignatiello, J.J. and White, E.D. (2015) "Effect of heredity and sparsity on second-order screening design performance", *Quality and Reliability Engineering International*, 31(3), 355-368.

DuMouchel, W. and Jones, B. (1994). "A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model". *Technometrics*, 36(1), 37-47.

Edwards, D.J. and Mee, R.W. (2011) "Fractional Box-Behnken designs for one-step response surface methodology". *Journal of Quality Technology*, 43(4), 288-306.

Erler, A., de Mas, N., Ramsey, P., and Henderson, G. (2012) "Efficient biological process characterization by definitive-screening designs: the formaldehyde treatment of a therapeutic protein as a case study". *Biotechnology Letters*, 35(3), 323 – 329.

Errore, A., Jones, B., Li, W. and Nachtsheim, C.J. (2017) "Using Definitive Screening Designs to Identify active first- and second-order factor effects". *Journal of Quality Technology*. In press

Jones, B. and Nachtsheim, C.J. (2011) "A Class of three-level designs for definitive screening in the presence of second-order effects", *Journal of Quality Technology,* 43, 1-15.

Jones, B. and Nachtsheim, C.J. (2013) “Definitive screening designs with added two-level categorical factors”, *Journal of Quality Technology,* 45, 121-129.

Jones, B. and Nachtsheim, C.J. (2017) “Effective Design-Based Model Selection for Definitive Screening Designs”, *Technometrics*, In Press.

Hamada, M. and Wu, C.J., 1992. “Analysis of designed experiments with complex aliasing”, *Journal of Quality Technology, 24*(3), 30-37.

Hecht, E.S., McCord, J.P., and Muddiman, D.C. (2015) “*Definitive* screening design optimization of mass spectrometery parameters for sensitive comparison of filter and SPE purified, INLIGHT plasma Nglycans”. *Analytical Chemistry.* DOI: 10.1021/acs.analchem.5b01609 Publication Date (Web) 18: Jun 2015

Kirshnamoorthy, A., Montgomery, D.C., Jones, B.J. and Borror, C.M. (2015) “Analyzing no-confounding designs using the Dantzig selector” *International Journal of Experimental Design and Process Optimisation*”, Vol. 4, pp. 183-205.

Li, X., Sudarsnam, N. and Frey, D. (2006) “Regularities in data from factorial experiments”, *Complexity*, 11, 32-45.

Libbrecht, W., Deruyck, F. Poleman, H., Verberckmoes, A., Thybaut, J., De Clercq, J. and Van Der Voort, P. (2015). “Optimization of soft template mesoporus carbon synthesis using Definitive Screening Design”. *Chemical Engineering Journal*, 259, 126-134.

Marley, C. and Woods, D.C. (2010) “A comparison of design and model selection methods for supersaturated experiments”. *Computational Statistics and Data Analysis*, 54, 3158-3167.

Montgomery, D.C., Myers, R.W., Carter, W.H. and Vining, G.G. (2005) “The hierarchy principle in designed industrial experiments”. *Quality and Reliability Engineering International.* 21, 197-201.

Phoa, F., Pan, Y.H., and Xu, H. (2009) “Analysis of supersaturated designs via the Dantzig selector”. *Journal of Statistical Planning and Inference*, 139, 2362-2372.

Shmueli, G. (2010). “To Explain or to Predict”. *Statistical Science.* 25, 289-310.

Ramsey, P. and Ornek, D. (2013) “Characterization of a Biomanufacturing Fermentation Process Using Definitive Screening Designs”. JMP Discovery Summit Conference, Cary, NC.

Ramsey, P. and Yeung, E. (2015) “Applications of Designed Experiments in the Development of an Analytical Method for Glycoprofiling.” Society of Experimental Biology, Annual Conference, Prague, Czech Republic.

Renzi, P., Kronig, C., Carlone, A. Eröküz, Berkessel, A. and Bella, M. (2014) “Kinetic Resolution of Oxazinones: Rational exploration of chemical space through the design of experiments.” *Chemistry European Journal*. 20, 11768-11775.

Xiao, L., Lin, K.J.D., and Bai, F. (2012) “Constructing definitive screening designs using conferences matrices”, *Journal of Quality Technology*, 44, 2-8.

Weese, M.L., Smucker, B.J., and Edwards, D.J. (2015) “Searching for powerful supersaturated designs”. *Journal of Quality Technology*, 47(1), 66-84.