Response Surface Experiments: A Meta-Analysis

Rebecca A. Ockuly[a], Maria L. Weese[a,*], Byran J. Smucker[a], David J. Edwards[b], Le Chang[a]

[a] *Miami University, Oxford, OH*

[b] *Virginia Commonwealth University, Richmond, VA*

Abstract

Response Surface Methodology is a set of experimental design techniques for system and process optimization that is commonly employed as a tool in chemometrics. In the last twenty years, thousands of studies involving response surface experiments have been published. The goal of the present work is to study regularities observed among factor effects in these experiments. Using the Web of Science Application Program Interface, we searched for journal articles associated with response surface studies and extracted over 20,000 records from all Science Citation Index and Social Science Citation Index disciplines between 1990 and the end of 2014. We took a random sample of these papers, stratified by the number of factors, and ended up with a total of 129 experiments and 183 response variables. Extracting the data from each publication, we reanalyzed the experiments and combined the results together in a meta-analysis to reveal information about effect sparsity, heredity, and hierarchy. We empirically quantify these principles to provide a better understanding of response surface experiments, to calibrate experimenter expectations, and to guide researchers toward more realistic simulation scenarios and improved design construction.

Keywords: Box-Behnken Designs; Central Composite Designs; Effect Heredity; Effect Hierarchy; Effect Sparsity;

1. Introduction

Using a statistically designed experiment in industry is typically done for one of two reasons: (1) to determine the most important factors related to the response; or (2) to estimate the factor settings that optimize a response. The latter objective is the goal of response surface methodology (RSM) [4, 21]. This widely influential process improvement strategy began in the chemical industries and continues to play a prominent role in a wide variety of chemical and other scientific disciplines, including chemistry, biotechnology,

[*] Corresponding author
Email address: weeseml@miamioh.edu (Maria L. Weese)

environmental science, food science, and biochemistry. The focus of this work is to better understand the general methodology by reanalyzing a large number of published RSM experiments from across these areas.

For response surface experiments that include several quantitative factors and a single response, an assumption is made that the response surface representing the system under study is adequately approximated by a second-order polynomial. Designs used for this purpose allow for enough experimental runs to fit the second-order model and predict the factor levels that produce the optimum response. Because prediction is inherent to this process, analysts should consider which terms to retain from the second-order model [20]. If one can observe patterns across many performed response surface experiments, experimenter expectations can be refined, research efforts can exploit these patterns to construct better design and analysis techniques, and guidance is provided for investigations that require simulations from response surface models. In this article, we investigate three such patterns—called regularities by Li et al.—that have been formalized in the context of two-level, factorial-type experiments as effect *sparsity*, *heredity*, and *hierarchy* [26]. These regularities have not been extensively studied in the second-order response surface model.

There are several slices of the vast RSM literature that are informed by the results of this work. First, one-step response surface methods [6, 14, 11, 10, 8], including definitive screening designs [13], are becoming increasingly popular. The general approach leans quite heavily on effect sparsity. Specifically, in order to both screen and optimize using a single experiment, one must use a design that can fit a full quadratic model, whether for the full number of factors or for some projection into a subset of the factors. Our work quantifies the level of sparsity that is typical in standard response surface studies.

There are also several examples in the literature of studies or methods that require simulation from response surface models. Allen and Yu build small response surface designs and rely on certain beliefs about effect sparsity and the size of second- and third-order parameters. Their work is in the context of simulation optimization, which requires assumptions about the nature of the system being considered. There are also a few other works that simulate from response surface models for various reasons [19, 8, 5]. Recently, several investigations have involved simulating from main effects and/or interaction models in order to assess the effectiveness of designs and analysis methods [18, 9, 25], and we expect that this sort of empirical investigation is used more frequently to study second-order designs and analysis methods as well. We believe that such simulation studies can be made more realistic by utilizing the results of this work.

Our primary goal is to perform a meta-analysis of response surface studies to empirically study effect sparsity, heredity, and hierarchy. We collect experimental data and perform a meta-analysis on 183 experimental responses from 129 published papers downloaded from the Web of Science database [24]. In what follows, we provide some additional background on response surface methodology as well as the regularities we studied, followed in Section 3 by an example that illustrates the nature of our approach. Section 4 provides a discussion of how the RSM studies were collected and sampled. Results of the meta-analysis are provided in Section 5 and Section 6 concludes the article with a summary and discussion.

2. Background

2.1. Response Surface Methodology

Response surface methods (RSM) are employed when attempting to find the factor level combinations that optimize the response(s) in an experiment. Often, a screening experiment is initially performed in order to reduce the number of factors upon which to subsequently experiment. Following a screening experiment, the classical response surface methodology prescribes a small design, such as a two-level fractional factorial, from which a first-order model is fit and rapid improvement attained via steepest ascent/descent [21, p. 234]. This procedure may iterate several times before curvature is detected, at which point a second-order design, such as a central composite [4] or a Box-Behnken [3] among other possibilities, is used to optimize the response via a full second-order model which includes all main effects, two-factor interactions, and quadratic terms. The second-order model, in k factors, is given as:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} X_i X_j + \epsilon,$$

with standard independence and normality assumptions on ϵ.

In this work, we focus on published experiments that employ the final RSM step in the form of either a central composite (CCD) or Box-Behnken design (BBD), given their popularity in practice. Figure 1 shows examples of a three-factor CCD and a three-factor BBD. More details regarding these design types are found in Myers et al. or Wu and Hamada.

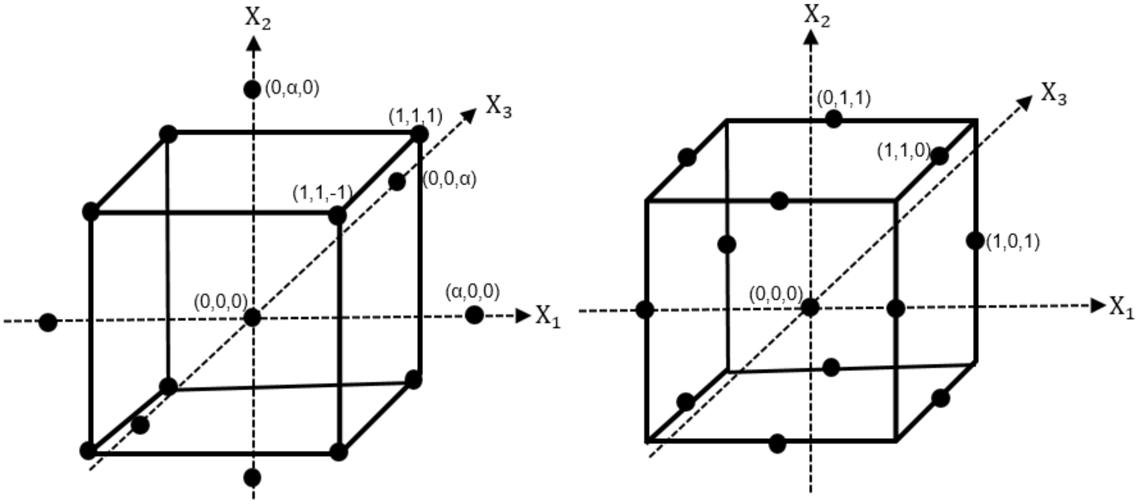

Figure 1: Examples of a 3-factor Central Composite Design (left) and a 3-factor Box-Behnken Design (right).

2.2. Regularities in Response Surface Experiments

Li et al. performed an empirical analysis of 46 two-level full factorial experiments, which yielded a total of 113 responses. They reanalyzed these experiments and the resulting meta-analysis produced quantifications of effect sparsity, heredity and hierarchy in the context of factorial experiments. Here, we broaden consideration of these regularities to second-order experiments. In what follows, we refer to *active effects* or *active terms* as those that are large enough to retain using an appropriate statistical criterion that we specify in Section 3.

Traditionally, a factorial experiment exhibits *effect sparsity* if only a small proportion (say, 20% or fewer) of the terms in the factorial model are active. In the response surface context, we extend effect sparsity to mean that only a small proportion of terms in the full second-order model are active. We consider overall effect sparsity, but also consider it separately for main effects, two-factor interactions, and quadratic terms. Note that effect sparsity is not the same as *factor sparsity*, which assumes that effect activity occurs among only a few of the factors. The effect sparsity assumption may hold even if factor sparsity does not.

The *effect heredity* principle assumes that the activity of an effect depends on its parent effects [26]. Under *strong effect heredity*, a two-factor interaction may be active only if both of its parent main effects are active, whereas *weak effect heredity* allows an active two-factor interaction when at least one parent is active [7]. In RSM, we preserve this definition with respect to two-factor interactions, but broaden the notion to encompass relationships between quadratic terms and both main effects and interactions. These details are described in Section 3. Note that effect heredity is not the same thing as functional marginality [22, 23]. While heredity describes the results of an analysis of experimental data, marginality is a philosophical inclusion of parent terms when a child is included. In other words, functional marginality forces models containing higher-order

Table 1: Three-factor Box-Behnken design for one response from Garrote et al.

Concentration	Temperature	Time	NaOH Penetration Thickness (mm)	tan(NaOH Penetration Thickness)
-1	-1	0	0.00	0.00
-1	0	-1	0.00	0.00
0	-1	-1	0.05	0.05
-1	0	1	0.10	0.10
1	0	-1	0.27	0.28
0	1	-1	0.27	0.28
1	-1	0	0.62	0.71
-1	1	0	0.62	0.71
0	-1	1	0.96	1.43
0	0	0	1.22	2.73
0	0	0	1.24	2.91
0	0	0	1.24	2.91
1	1	0	2.16	-1.49
1	0	1	2.67	-0.51
0	1	1	2.84	-0.31

terms to include all lower-order parent terms regardless of their activity. For instance, if a model includes the three-factor interaction ABC, marginality would require that all of the parent main effects (A, B, C) and two-factor interactions (AB, AC, BC) must also be included in the model.

In factorial experiments, *effect hierarchy* asserts itself when lower-order terms, such as main effects, are larger than higher-order terms, such as two-factor or larger interactions. A consequence of this is that lower-order terms are more likely to be active than higher-order terms. In the RSM context, effect hierarchy refers to comparisons between the sizes of main effects and the sizes of two-factor interactions and quadratic effects.

3. Example

We now present an example analysis of a response surface experiment to illustrate the regularities under investigation. Garrote et al. describe a one-stage experiment that studies the chemical peeling of potatoes. They use a three-factor BBD (with three center point runs) to study the effects of concentration (A), temperature (B), and time (C) on two responses, heat-ring formation and sodium hydroxide (NaOH) penetration. The goal of experimentation is to determine optimal processing conditions. The authors do not mention the use of a screening design or steepest accent prior to running the BBD nor do they discuss their choice of the levels for each factor. A potential downside of foregoing these initial experimental steps is an increased risk that the region of experimentation does not contain the optimum. Here we present the analysis of the NaOH penetration response. Table 1 displays the coded factors as well as the response for the 15 treatment combinations in the potato peeling experiment.

Table 2: Results of the analysis of the BBD with the transformed response from Garrote et al., $R^2 = 0.981$ and $R^2_{\text{adj}} = 0.946$.

	Estimate	Std. Error	t-statistic	p-value	FDR p-value
A	-0.22871	0.10759	-2.1257	0.08687	0.11169
B	-0.3761	0.10759	-3.4957	0.01736	0.03125
C	0.01301	0.10759	0.1209	0.90846	0.90846
AB	-0.73098	0.15216	-4.8042	0.00487	0.01095
AC	-0.22177	0.15216	-1.4575	0.20477	0.23037
BC	-0.49154	0.15216	-3.2305	0.02319	0.03479
A^2	-1.63176	0.15837	-10.3036	0.00015	0.00133
B^2	-1.23752	0.15837	-7.8142	0.00055	0.00165
C^2	-1.25367	0.15837	-7.9162	0.00052	0.00165

In the published analysis of the NaOH penetration response, the authors noticed significant lack of fit and addressed this via a transformation of the response. After evaluating several choices, the tangent transformation was selected. As such, we also analyzed the transformed response, tan(NaOH penetration).

The following describes our methodology for assessing each regularity (sparsity, hierarchy and heredity) and any additional information we collected regarding the response surface studies we examined. Each design and response combination was reanalyzed using the `rsm` package in R [15]. Both ordinary unadjusted p-values and *false discovery rate*-adjusted p-values were examined to declare active effects, assuming the traditional $\alpha = 0.05$. The false discovery rate (FDR) [2] is the expected proportion of false discoveries amongst rejected hypotheses, and the associated adjustment is used to reduce problems associated with multiple comparisons.

Based on the criteria (unadjusted p-value or FDR p-value) and the corresponding effects declared active, we calculate the proportion of active main effects, two-factor interactions, and quadratic effects as a means of assessing effect sparsity. For the experiment shown in Table 1, Table 2 gives the estimates, t-statistics, unadjusted p-values and FDR p-values for the analysis of the transformed response. Using a cutoff of $\alpha = 0.05$, one would consider B, AB, BC, A^2, B^2, C^2 to be active using both unadjusted p-values and FDR p-values. Notice that in this particular example, adjusting the p-values for the false discovery rate does not change the conclusions, but this is certainly not always the case. The proportion of active main effects is calculated as $(1/3) * 100\% = 33\%$ using both p-value criteria. Likewise, $(2/3) * 100\% = 67\%$ and $(3/3) * 100\% = 100\%$ of the two-factor interactions and quadratic effects, respectively, are declared active.

To assess effect hierarchy, we use an effect size measured as the absolute value of the t-statistic for each effect. For each effect type, including effects deemed inactive via significance testing, we report the average and median of the t-statistics (see Table 3). In this example, quadratic effects clearly dominate followed by two-factor interactions and main effects.

Table 3: Example estimated average and median effect sizes for the different effect types.

Effect Type	Mean	Median
Main Effects	1.914	2.126
Two-Factor Interactions	3.164	3.231
Quadratic Effects	8.678	7.916

To investigate heredity, we classify active two-factor interactions as exhibiting strong effect heredity, weak effect heredity, or no effect heredity. A strong heredity interaction is defined as the event $\{AB \text{ active} \mid A \text{ and } B \text{ active}\}$, weak heredity as $\{AB \text{ active} \mid \text{one of } A \text{ or } B \text{ active}\}$, and no heredity as $\{AB \text{ active} \mid A \text{ and } B \text{ inactive}\}$. Using the results in Table 2, only one main effect is active (for both unadjusted and FDR p-values). Thus, in this example, is is not possible for a two-factor interaction to exhibit strong heredity. Since there is one active main effect, B, there are two possible weak heredity interactions: AB and BC. As both of these interactions are considered active, 100% of eligible weak heredity terms actually exhibit it. Finally, given the activity of the B main effect, only AC could exhibit no effect heredity. However, since this term is deemed inactive, none of the two-factor interactions display no heredity.

In addition to the traditional two-factor interaction heredity, one may also require that in order for a quadratic effect to be active, its corresponding main effect must be active [27]. As a consequence, define main effect/quadratic heredity as $\{A^2 \text{ active} \mid A \text{ active}\}$, and no main effect/quadratic heredity as $\{A^2 \text{ active} \mid A \text{ inactive}\}$. From Table 2, we see that all three quadratic effects are deemed active. Since B is the only active main effect, however, the only quadratic effect eligible to exhibit main effect/quadratic heredity is B^2. As B^2 is indeed active, 100% of the quadratic effects eligible to exhibit main effect/quadratic heredity are active. Similarly, both A^2 and C^2 are eligible to exhibit no main effect/quadratic heredity since neither A nor C main effects are active. Thus, 100% of the quadratic effects eligible to exhibit no main effect/quadratic heredity are active.

Despite the fact that a quadratic effect and a two-factor interaction are of the same order (i.e., second order effects), preliminary investigations suggested that quadratic effects tended to be larger than two-factor interactions. As a consequence, we also investigate a heredity-like relationship among two-factor interactions and quadratic effects. Strong quadratic/interaction heredity is defined as $\{AB \text{ active} \mid A^2 \text{ and } B^2 \text{ active}\}$, weak quadratic/interaction heredity as $\{AB \text{ active} \mid \text{one of } A^2 \text{ or } B^2 \text{ active}\}$, and no quadratic/interaction heredity as $\{AB \text{ active} \mid A^2 \text{ and } B^2 \text{ inactive}\}$. Given that all three quadratic effects in our example are declared active by both criteria, only strong quadratic/interaction heredity is possible. There are a total of three possible strong quadratic/interactions and two of them (AB and BC) are deemed active. Thus, two-thirds of the two-factor interactions eligible to exhibit strong quadratic/interaction heredity are active.

Table 4: Additional analysis output from the Garrote et al. experiment.

	df	Sum Sq	Mean Sq	F value	Pr(>F)
Linear effects	3	1.5514	0.5171	5.5845	0.0471534
Two factor interactions	3	3.3005	1.1002	11.8802	0.0103284
Quadratic effects	3	18.5301	6.1767	66.6998	0.0001874
Residuals	5	0.463	0.0926		
Lack of Fit	3	0.4416	0.1472	13.7559	0.0685281
Pure error	2	0.0214	0.0107		
Stationary Point: (-0.0395, -0.1478, 0.0377) Eigenvalues: (-0.9311, -1.2662, -1.9257)					

Table 4 displays additional output from the analysis of the chemical peeling experiment using the `rsm` package. The lack-of-fit test, when analyzing the transformed response, tan(NaOH Penetration), is not significant at the 0.05 level ($p = 0.07$). We also note that the stationary point is a maximum and located inside the design region. The additional information collected in Table 4 is recorded for each experiment in our meta-analysis.

4. Data Collection and Sample Considerations

The population of interest in this study is the set of published response surface studies that used the traditional central composite or Box-Behnken designs. As such, we performed a search using the Web of Science (WoS) Application Program Interface [24]. We limited our consideration to papers published between January 1, 1990 and December 31, 2014, and included "central composite" and "Box-Behnken" along with "Response Surface" as search terms. A total of 24,286 papers were returned when performing this general search, and since most response surface designs are fairly small in terms of the number of factors ($k = 2, 3, \text{or } 4$ factors) we oversampled for larger experiments ($k = 5, 6, \text{and } 7$). To do this, we searched from the full population of papers on the terms "five factors", "six factors", and "seven factors" and found 119, 22, and 21 papers, respectively.

To obtain an initial sample of papers, we sampled all 21 in the seven-factor category, all 22 from the six-factor, and 75% of those papers from the "five factors" search. In addition, 188 papers from the rest of the population were sampled, for a total of 320 papers. From this original sample, many papers were disqualified for various reasons and ultimately, we used 129 papers (Table 5). In order to accurately infer to the population of published experiments, we weighted the sample statistics according to our stratified sampling procedure. See Appendix A for details regarding the sampling procedure and the calculation of the sampling weights.

In Section 3, we presented an example that contained multiple responses for a single BBD. Ideally, one could treat these multiple responses as independent, but it is not uncommon for multiple responses from a single experiment to be correlated. Our subsequent meta-analysis includes the computation of standard errors, which are computed under the assumption that the observations are uncorrelated. To increase the plausibility of this assumption, we tested all pairwise correlations among responses in the same experiment for significant differences from zero (using p-values adjusted for multiple comparisons and $\alpha = 0.05$). If a correlation was significantly different from zero, we removed one of the responses in the correlated pair from the meta-analysis. This left us with a sample of 183 responses from 129 separate papers.

Upon analyzing the 183 responses as described in Section 3, we found 68 (37%) showing significant lack-of-fit (p-value < 0.05) upon fitting the full second order model. This suggests that a substantial proportion of published RSM experiments used error variance estimates that were biased. This raises a concern not only regarding the published experimental results, but also regarding our meta-analysis. We handled this by performing the meta-analysis on all 183 responses as well as the 115 that did not show any significant lack of fit. We found that considering only the responses showing no statistically significant lack-of-fit did not change the general conclusions and, as such, the results presented in the following sections are from a meta-analysis of all 183 responses. However, the interested reader is referred to Appendix B for results of the re-analysis of the 115 experiments showing no lack-of-fit.

5. Meta-Analysis Results

Out of the 129 papers from which we obtained experimental data, 101 of them were CCDs, and 28 were BBDs. Of these, 81 were given in coded units, and 48 were given in raw, uncoded units and had to be coded prior to analysis. Forty-six of the papers specifically mentioned the performance of a preliminary screening experiment, while the other 83 did not.

The 129 experiments ranged from 9 to 100 runs, with a median run size of 20 and a mode of 16. Of all the responses in the sample, most were associated with three-factor experiments; the full breakdown is given in Table 5. Note that these results reflect the stratified sampling we applied and so give an unfair impression, for instance, of the prevalence of five-factor experiments. The axial distances used in the CCDs ranged from 0.6 to 2.83, with a median axial distance of 1.69 and a mode of 2. Remarkably, not a single face-centered CCD was encountered in the sample. All but four experiments contained replicated runs.

Table 5: The distribution of the number of factors associated with each paper and response in our sample of experiments

Number of Factors	Papers	Responses
2	16	29
3	41	70
4	28	29
5	37	47
6	5	5
7	2	3

Figure 2 shows that the most represented field of study in the population (correcting for stratification), based on the SCIE/SSCI category of the publication, is Chemistry with an estimated 30% of papers belonging to that field; Biotechnology and Engineering follow closely with 29% and 22% of the papers, respectively.

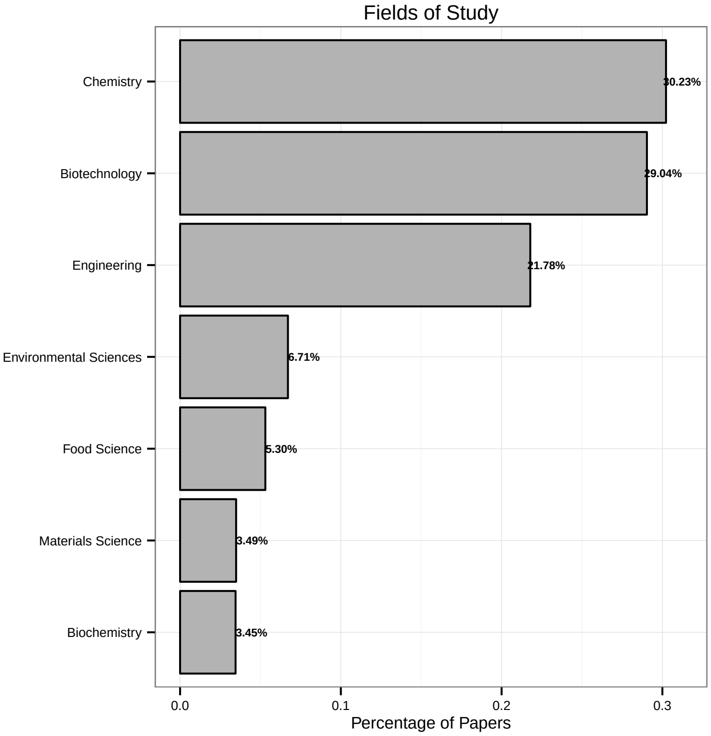

Figure 2: The fields of study represented

As more fully described in Section 3, we analyzed each of the 183 responses to quantify sparsity, hierarchy and heredity. The estimates reported in the rest of this section have been weighted to reflect the stratified sample (see Table A3 in the Appendix), unless otherwise noted. The `survey` package in R [17] was employed to obtain estimates and standard errors appropriately adjusted for stratification.

5.1. Effect Sparsity And Hierarchy

Table 6 displays the weighted proportions and standard errors of the active effect types (main effect, quadratic, two-factor interaction) using unadjusted and FDR adjusted p-values, respectively. Via unadjusted p-values, we found 55% of main effects to be active, compared to 20% of two-factor interactions and 47% of quadratic effects. The more conservative FDR-adjusted p-values show 48% of main effects, 16% of two-factor interactions, and 41% of quadratic effects as active. Li et al. reported smaller proportions of main effects (41%) and two-factor interactions (11%) as active in the context of two-level full factorial designs. We suspect the higher proportions found in the response surface studies are primarily due to formal or informal screening conducted prior to the second-order experiments. Overall, 39% (SE 0.03) of all factor effects were deemed to be active (34% (SE 0.03) using FDR-adjusted p-values).

Table 6: Effect sparsity, measured as the proportion of active effect types.

Effect Type	Unadjusted p-values (SE)	FDR p-values (SE)
Main Effects	0.55 (0.03)	0.48 (0.04)
Two-Factor Interactions	0.20 (0.03)	0.16 (0.03)
Quadratic Effects	0.47 (0.04)	0.41 (0.04)

Tables 7 and 8 provide additional details on active effect types by the number of experimental factors. Note these proportions and standard errors are not weighted for stratification. Overall, as in Table 6, we find that main effects are most likely to be active, followed by quadratic effects and then two-factor interactions. Although quadratic effects and two-factor interactions are of the same order, it is interesting to note the preponderance of active quadratic effects over two-factor interactions. This observation may indeed lend itself to the heredity-like relationship among quadratic effects and two-factor interactions mentioned in Section 3 and investigated later in this section.

Table 7: Proportion of active effect types by k for the sample, using unadjusted p-values. Note that these proportions and standard errors are not calculated using the stratified sampling weights.

k	# Experiments	Main Effects (SE)	Two-Factor Interactions (SE)	Quadratic Terms (SE)
2	29	0.45 (0.07)	0.14 (0.06)	0.34 (0.06)
3	70	0.49 (0.03)	0.20 (0.03)	0.44 (0.03)
4	29	0.70 (0.04)	0.22 (0.03)	0.52 (0.05)
5	47	0.54 (0.03)	0.22 (0.02)	0.46 (0.03)
6	5	0.70 (0.08)	0.29 (0.05)	0.63 (0.09)
7	3	0.62 (0.11)	0.29 (0.06)	0.38 (0.11)

Table 8: Proportion of active effect types by k for the sample, using FDR-adjusted p-values. Note that these proportions and standard errors do not use the stratified sampling weights.

k	# Experiments	Main Effects (SE)	Two-Factor Interactions (SE)	Quadratic Terms (SE)
2	29	0.33 (0.06)	0.14 (0.06)	0.26 (0.06)
3	70	0.41 (0.03)	0.15 (0.02)	0.36 (0.03)
4	29	0.65 (0.04)	0.17 (0.03)	0.49 (0.05)
5	47	0.40 (0.03)	0.16 (0.02)	0.33 (0.03)
6	5	0.66 (0.09)	0.27 (0.05)	0.63 (0.09)
7	3	0.52 (0.11)	0.24 (0.05)	0.29 (0.10)

As stated previously, effect hierarchy is quantified by using the absolute value of the t-statistics (i.e., $|t| = |\hat{\beta}_i/\mathrm{SE}_{\hat{\beta}_i}|$) as a measure of effect size. Table 9 shows these mean effect sizes, along with their standard errors. The means are inflated by particularly large estimates that constituted about 10% of the sampled papers. Since no reason was identified to exclude them from the sample we retained them, but include median estimates of effect sizes as well (Table 9). From the results, we see a rough rule of thumb: main effects are about 1.25 times as large as quadratic effects, which are about twice as large as two-factor interaction effects. This is similar to Li et al., who reported the median main effect strength to be about four times that of two-factor interactions for factorial experiments. Figures 3 and 4 graphically show distributions of effect sizes.

Table 9: Estimated average and median effect sizes for the different effect types, weighted to account for stratified sample.

Effect Type	Mean (SE)	Median (SE)
Main Effects	5.48 (0.41)	2.73 (0.33)
Two-Factor Interactions	1.60 (0.11)	0.95 (0.08)
Quadratic Effects	4.00 (0.25)	2.12 (0.21)

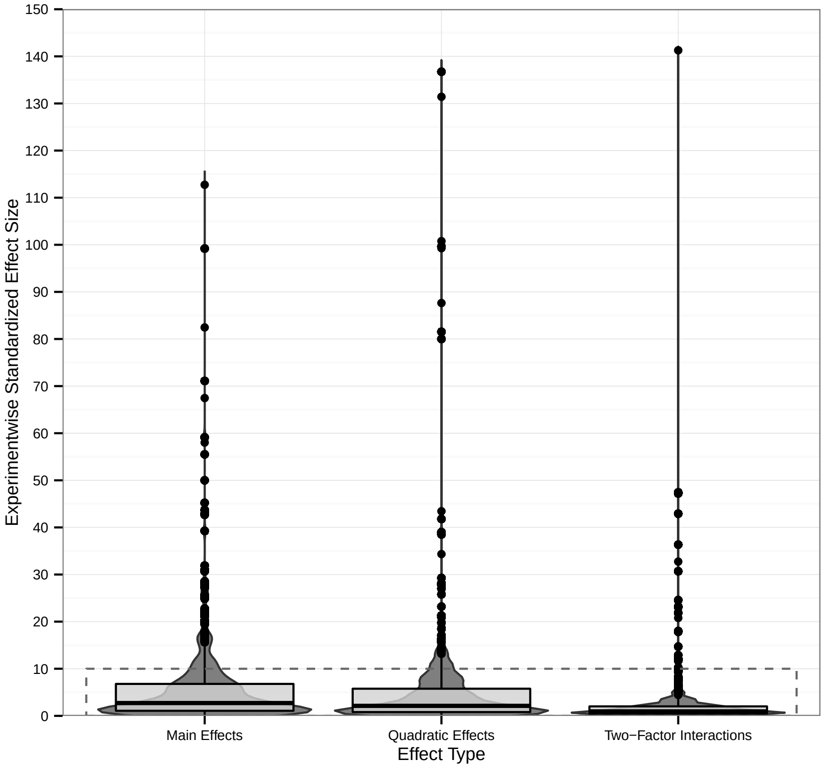

Figure 3: Boxplots and kernel density plots of standardized effect sizes, weighted to reflect the sampling scheme. The numerous outliers make it difficult to compare across effect types. The part of the plot within the dashed line box is examined in Figure 4.

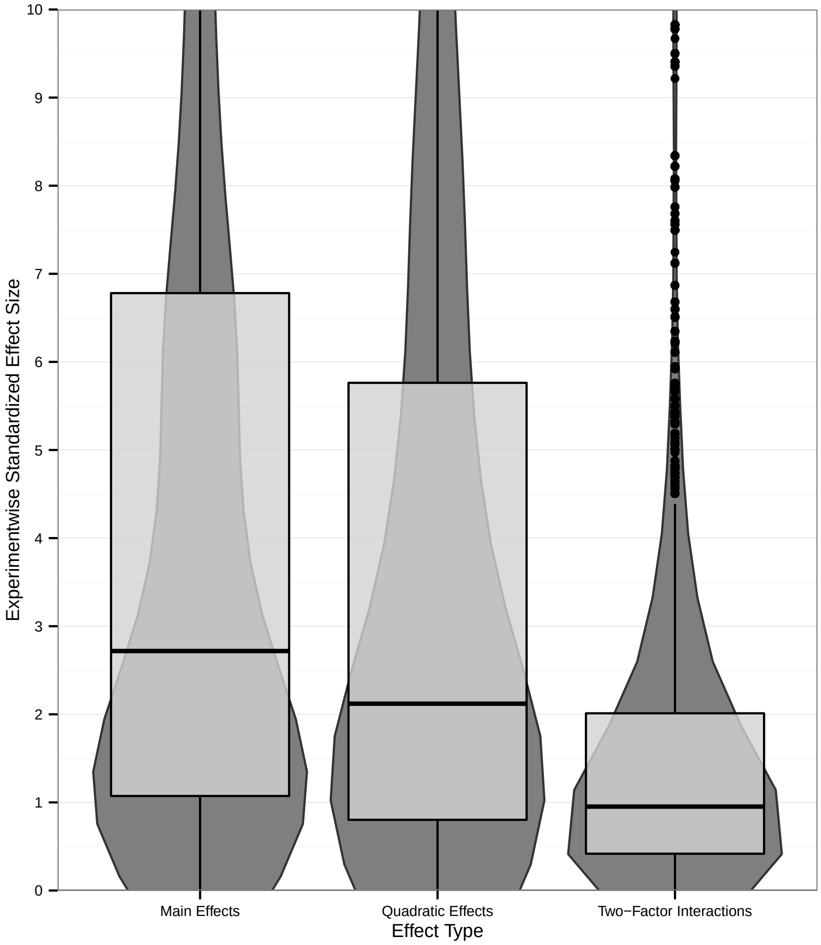

Figure 4: Boxplots with kernel density plots for standardized effect sizes, weighted to reflect the sampling scheme, using rescaled y-axis. The middle line indicates the median.

5.2. Heredity

Table 10 shows the proportions and standard errors of interactions exhibiting strong, weak or no heredity, weighted for stratification. Using unadjusted p-values and the definitions given in Section 3, the probability that a two-factor interaction is active given that both of its parent main effects are active is estimated to be 33%. This probability decreases to 16% if only one of the parents is active and to 7% if neither are active. The proportions are slightly less using the FDR-adjusted p-values. Overall, we find clear support of the traditional view that two-factor interactions are more likely to be active if one or more of its parent main effects are active. Furthermore, an interaction obeying strong effect heredity is twice as likely to occur than one obeying weak heredity.

Table 10: Traditional heredity, measured as the proportion of active two-factor interactions when 0, 1, or 2 parent main effects are active.

Strength	Unadjusted p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong	0.33 (0.04)	0.32 (0.05)	$A, B,$ and AB
Weak	0.16 (0.04)	0.13 (0.04)	A and AB
None	0.07 (0.03)	0.04 (0.02)	AB

As described in Section 3, we also examined the heredity relationship between main effects and quadratic effects. Table 11 displays the proportions and standard errors of active quadratic effects that do and do not exhibit main effect-quadratic heredity using unadjusted and FDR-adjusted p-values. As expected, a quadratic effect is more likely to be active if its corresponding main effect is active. Perhaps more surprising is that quadratic effects are so often active without their parents (33% unadjusted; 28% FDR-adjusted). This may be more likely to happen when the stationary point is located within the design region (see Section 5.3 for additional discussion).

Table 11: Main effect/quadratic heredity, measured as the proportion of active quadratic effects when 0 or 1 of its parent main effects are active.

Heredity	Unadjusted p-values (SE)	FDR p-values (SE)	Example of Active Terms
Quadratic Heredity	0.58 (0.04)	0.55 (0.05)	A and A^2
No Quadratic Heredity	0.33 (0.05)	0.28 (0.05)	A^2

The concept of heredity among quadratic effects and two-factor interactions may also be of practical interest, since designs that are becoming more common—such as algorithmically generated optimal designs and Definitive Screening Designs [13]—may exhibit correlation between quadratic effects and two-factor interactions. Table 12 shows the proportions and standard errors of interactions exhibiting strong, weak and no quadratic/interaction heredity for unadjusted p-values and FDR-adjusted p-values. Notably, the strength of heredity here is similar to the standard heredities given in Table 10, and in fact the weak and none versions of the quadratic/interaction heredity are more prevalent.

Table 12: Quadratic-interaction heredity, measured as the proportion of active two-factor interactions when 0, 1, or 2 of its associated quadratic terms are active.

Strength	Unadjusted p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong	0.30 (0.05)	0.27 (0.05)	$A^2, B^2,$ and AB
Weak	0.23 (0.05)	0.21 (0.05)	A^2 and AB
None	0.09 (0.03)	0.05 (0.02)	AB

Finally, we combined the results of traditional main effect-interaction heredity and quadratic-interaction heredity to create a so-called "expanded" heredity (Table 13). The first listed strength refers to the heredity relationship between the active main effects and two-factor interactions. The second strength refers to the heredity relationship between the active quadratic effects and two-factor interactions. For example, the strength labeled "Strong-Strong" is the probability a two-factor interaction is active given that both of the corresponding main effects and both of the corresponding quadratic effects are active. Likewise, "Weak-Weak" indicates the probability a two-factor interaction is active given that one of the involved main effects is active and one of the involved quadratic main effects are active. Note this does not imply the active main effect and active quadratic main effect involve the same factor. We see that a two-factor interaction is most likely to be active when both parent main effects and associated quadratic effects are active.

Table 13: Expanded heredity, measured as the proportion of active interactions with some combination of active main effects and quadratic effects.

Strength	Unadjusted p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong-Strong	0.40 (0.06)	0.38 (0.07)	$A, B, A^2, B^2,$ and AB
Strong-Weak	0.36 (0.08)	0.37 (0.09)	$A, B, A^2,$ and AB
Strong-None	0.16 (0.06)	0.16 (0.06)	$A, B,$ and AB
Weak-Strong	0.26 (0.08)	0.22 (0.08)	$A, A^2, B^2,$ and AB
Weak-Weak	0.15 (0.05)	0.12 (0.05)	$A, A^2,$ and AB or $A, B^2,$ and AB
Weak-None	0.08 (0.04)	0.07 (0.04)	A and AB
None-Strong	0.07 (0.06)	0.10 (0.07)	$A^2, B^2,$ and AB
None-Weak	0.08 (0.06)	0.07 (0.06)	A^2 and AB
None-None	0.06 (0.03)	0.01 (0.01)	AB

5.3. Stationary Points

A classical analysis of the second-order model includes an estimate of the location of the response surface stationary point. Such a stationary point may represent a maximum, a minimum, or a saddle point; the latter occurs when the eigenvalues of the relevant matrix involving second-order regression coefficients have mixed signs [see 21]. We performed the stationary point analysis using the full model, including non-significant terms. Of the 183 responses we analyzed, 102 had stationary points occurring within the design region, and 81 occurring outside the design. Adjusting for the stratification of the sample, we estimate that 56.7% of RSM experiments in the population had a stationary point inside the design region. Of the stationary points we found, both inside and outside the experimental region, the weighted proportions and standard errors for maxima, minima, and saddle are 39.8% (0.04), 5.5% (0.02), and 54.7% (0.04), respectively. Interestingly, when we eliminated the experiments containing significant lack of fit, we estimate 58.8% had a stationary point inside the design region and the weighted proportions for maxima, minima, and saddle are 40.6% (0.06), 5.3% (0.03), and 54.1% (0.06) respectively, which are relatively unchanged.

The original conception of response surface methodology by Box and Wilson was an iterative process that culminated in a second-order design to capture curvature effects in the interior of the design space. In such a case, second-order effects would dominate. Based on a reviewer's suggestion, we investigated this by performing the analysis of Sections 5.1 and 5.2 separately for those experiments for which the stationary point was inside the design region and for those outside. The same sampling weights are used as those used previously. We find that when the stationary point is inside the design region, quadratic effects are more common (Table 14) and much larger (Table 15).

Table 14: Effect sparsity, measured as the proportion of active effect types and weighted to account for stratified sampling, for responses with stationary points inside, (a), and outside, (b), the design region.

(a) Inside

Effect Type	Unadjusted p-values (SE)	FDR p-values (SE)
Main Effects	0.50 (0.04)	0.44 (0.05)
Two-Factor Interactions	0.25 (0.04)	0.21 (0.04)
Quadratic Effects	0.60 (0.05)	0.54 (0.06)

(b) Outside

Effect Type	Unadjusted p-values (SE)	FDR p-values (SE)
Main Effects	0.61 (0.05)	0.54 (0.05)
Two-Factor Interactions	0.15 (0.03)	0.10 (0.02)
Quadratic Effects	0.31 (0.05)	0.26 (0.05)

Table 15: Estimated average and median effect sizes for the different effect types, weighted to account for stratified sample, for responses with stationary points inside, (a), and outside, (b), the design region.

(a) Inside

Effect Type	Mean (SE)	Median (SE)
Main Effects	5.16 (0.56)	2.30 (0.26)
Two-Factor Interactions	1.91 (0.19)	1.04 (0.14)
Quadratic Effects	5.36 (0.41)	3.69 (0.54)

(b) Outside

Effect Type	Mean (SE)	Median (SE)
Main Effects	5.88 (0.60)	4.09 (0.77)
Two-Factor Interactions	1.28 (0.08)	0.95 (0.09)
Quadratic Effects	2.39 (0.24)	1.25 (0.22)

6. Discussion and Conclusions

6.1. Discussion

As described earlier in this article, we oversampled for large response surface designs, and subsequently examined *all* six- and seven-factor designs that arose in our keyword search. This produced some interesting - and perhaps concerning - results. For instance, we found two separate articles, from the same authors, describing what appears to be the same design (same factors and response variables), though the response values were different. A generous assumption is that the authors ran two separate experiments and published them separately, perhaps highlighting different aspects of the experiments. This was not the only curious case we found; in all, three pairs of papers exhibited this sort of similarity. What is more, we also found another pair of papers that shared a common response variable as well as the same response values. We omitted one of these response variables from our meta-analysis but retained the others.

There are several cautions we wish to highlight regarding our work. First, we obtained 129 experiments which initially resulted in a total of 262 responses. As described in Section 4, the number of responses was reduced to 183 by removing highly correlated responses from each experiment. Thus, in our analysis, we treated each response as independent observational units. A more conservative approach would have been to choose just one response from each paper. However, this would have required us to discard a lot of information. We believe a reasonable compromise was to omit those responses that were flagged as highly correlated with another response in a particular experiment. Secondly, the population to which we can generalize is limited to the published scientific literature between 1990 and the end of 2014. There are many other experiments that have been conducted in industry, or may have been run but not published because results were uninteresting or for some other reason. It is possible that these experiments exhibit different characteristics than those studied here.

One reviewer suggested that it may also be interesting to note differences in regularities among research fields. Although we do not suspect that any formal conclusions can be reached regarding field differences with our current data (as we did not conduct stratified sampling by field and, as a result, some fields are represented by only a very small number of experiments), we can make a few empirical statements regarding the top three research fields (Chemistry: 49 experiments, Biotechnology: 41 experiments, and Engineering: 32 experiments). Overall, sparsity is consistent across these three fields and the rule of thumb for hierarchy (Section 5.1) appears to hold. Strong main effect/interaction heredity dominates all three of these fields. In contrast, chemistry exhibited the largest proportion of active quadratic effects with no corresponding active main effect. Furthermore, the stationary point analysis for Biotechnology revealed almost three quarters being saddle points and over half of these stationary points outside the experimental region. Reasons for such differences among fields is a worthy topic for future research.

6.2. Conclusions

In this article, we have reported an extensive meta-analysis of response surface studies from the literature over the last twenty-five years. In particular, we have investigated effect sparsity, hierarchy, and heredity in the context of second-order experiments. These experiments are most commonly executed in the chemical and related disciplines.

Despite the caveats of the previous section, we have learned several things. First, effect sparsity exists for response surface studies. Main effects and quadratic effects are markedly more common than two-factor interactions. The levels of sparsity are less pronounced than for factorial experiments, likely because of prior factor screening. Furthermore, quadratic effects exhibit a level of sparsity between interactions and main effects. Effect hierarchy results are quite similar to those discovered by Li et al.. A rough rule is that quadratic effects are twice the size of two-factor interactions, and main effects are 1.25 times the size of quadratics.

We also confirmed that several different types of heredity appear present in this population of studies. In particular, strong heredity is quite prominent, both when considering interactions and main effects, but also when examining quadratics and interactions. The findings regarding main effect-quadratic heredity were perhaps the most surprising. Roughly one-third of the time, when the main effect was inactive, the corresponding quadratic effect was found to be active. Over half of the time, when a main effect was active its corresponding quadratic effect was active.

We found that for those responses with stationary points inside the design region, that quadratic effects were much more common and the size of these effects were much larger, larger even than main effects. Based on the number of responses with stationary points outside the design region and the number of saddle points, it appears that the textbook situation—which includes an optima within the design region—is not common.

The conclusions from this study are useful to provide guidance for the construction of second-order designs. As a possible example, consider the conclusions regarding hierarchy and heredity. We found the magnitude of effects ordered largest to smallest as main effects, quadratics, then interactions, and also that it is not uncommon to observe an active quadratic effect with no corresponding active main effect. Perhaps, then, in the construction of a design where the goal is to estimate a second-order model, it would be prudent to distribute the correlation among effect columns to reflect this hierarchy and lack of heredity. Most often, the least amount of correlation is between the main effect columns, but perhaps placing slightly more correlation between the main effect columns in order to reduce the correlation between the interactions and quadratic columns can provide more power to detect those potentially smaller, but important, effects.

Many of the implications of this work are connected to recent work in one-step response surface designs (e.g. the Definite Screening Designs of Jones and Nachtsheim). For instance, we estimate that about 30% of the time, second-order terms do not correspond to an active main effect. Though these one-step RSM designs may be able to estimate full quadratic models in a subset of the active factors, our results suggest that quadratic effects often appear without connection to active main effects. Thus, analysis strategies that emphasize an initial main-effects only model may lead to misleading and/or ambiguous results.

Certainly, more could be said regarding the ramifications of this study on future design construction and analysis methods. We hope that these findings will serve to inform researchers and practitioners in the development and use of second-order response surface designs as well as serve as an impetus for further research.

Acknowledgments

We would like to thank Dr. Bradley Jones and Dr. Karsten Maurer for their helpful suggestions and advice as well as Mark Hermon for his assistance in accessing the WoS API.

Supplementary Materials

Meta-Analysis References: File containing a bibliography of all 129 articles used for the meta-analysis in this paper. (MetaBib.pdf)

Appendix A. Sampling Weight Details

As described in Section 4, our initial Web of Science search was quite general and included “central composite”, “Box-Behnken”, and “Response Surface” as search terms. A total of 24,286 papers were returned when performing this general search. Since most response surface designs are fairly small in terms of the number of factors ($k = 2, 3, \mathrm{or}\ 4$ factors) we oversampled for larger experiments ($k = 5, 6, \mathrm{and} 7$). To prepare for such a stratified sample, we performed subsequent searches adding the terms “five factors”, “six factors” and “seven factors”, respectively, to the previous search terms. These queries resulted in 119 (“five factors”), 22 (“six factors”), and 21 (“seven factors”) papers. The “Papers from WoS” column in Table A1 provides details of the initial WoS population, in terms of the strata of interest.

Table A1: Results of WoS search and stratified sample.

		Papers from WoS	Papers Sampled	Usable Papers	Approx. Popn. Size
Sampled Category	General	24,124	188	67	8,600
	Five	119	89	42	56
	Six	22	22	8	8
	Seven	21	21	12	12
	Total	24,286	320	129	8,676

From this initial population, we obtained a stratified random sample of papers to provide the data for meta-analysis. We sampled all 21 of the potential seven-factor experiments, all 22 of the potential six-factor experiments, 89 (75%) of the potential five-factor experiments, and 188 of the papers from the more general search. This resulted in an initial sample of 320 papers taken from the original 24,286. Individual inspection of the articles revealed that many of them were unacceptable for various reasons: they did not contain the experimental data; they were studying RSM instead of performing RSM; the designs were not standard CCDs or BBDs; etc. After removing such articles, we were left with 129 that were appropriate for reanalysis. These stratified sampling results, along with the suitable remnants, are given in the "Papers Sampled" and "Usable Papers" columns of Table A1. A complete bibliography of the 129 papers used for our meta-analysis is available in the supplementary materials.

For subsequent estimates regarding effect sparsity, heredity, and hierarchy, the stratification of the sample must be accounted for. That is, we must weight the estimates within strata by the probability that a particular sampling unit is in a particular stratum. This quest is complicated by several factors. First, the initial population of 24,286 is not the target population because many of the papers were unacceptable for reanalysis as described above. Thus, we must estimate the size of the target population which includes only suitable experiments and data available for reanalysis. Secondly, being sampled into the four categories (2-4, 5, 6, 7) is no guarantee that the data included in the sampled paper actually belonged in that category. The movement of the usable papers from sampled to actual categories is given in Table A2, based on an inspection of each paper. Thirdly, many papers had multiple responses for the same experiment, and we used results from all of the responses in our sparsity, hierarchy, and heredity estimates. Thus, we should account for this in the estimates of the weights.

Table A2: The number of papers sampled and actually in each stratum.

		Actual Category			
		Two–Four	Five	Six	Seven
Sampled Category	Two–Four	63	3	1	0
	Five	8	34	0	0
	Six	4	0	4	0
	Seven	10	0	0	2
	Total	85	37	5	2

To compute the weights, we first must estimate the size of each stratum in the true population of published RSM studies that used CCDs or BBDs. We do this by evaluating the proportion of initially sampled papers that were usable, and then using this proportion to estimate the number of papers in each stratum. For instance, there were 188 papers sampled from the "Two–Four" factors stratum and 67 of them were usable. Thus, our estimate of the size of this stratum in the WoS population is $(67/188) \cdot 24{,}124 \approx 8600$. Other estimates of this sort are given in Table A1.

Once these population sizes are estimated, they are normalized into probabilities (Table A3), and these are given in the “Weight (sampled)” column, so named because they are based upon the strata that the papers were initially sampled into. However, notice in Table A2 that being sampled into a stratum does not always mean the experiment is correctly categorized. The “Total” row of Table A2 contains the number of sampled papers correctly categorized. So, in order to estimate the probability of a sampling unit *actually* being in each strata, we can use Table A2 to compute probabilities of the sort $P(\text{actually in category } i|\text{sampled in category } j)$ and these are used along with the “Weight (sampled)” probabilities and the law of total probability to compute the probabilities of actually being in a particular stratum. These weights are given in the “Weight (actual)” column of Table A3. Notice that the “Two–Four” category has a lower weight and the “Five” and “Six” categories are upweighted, because a few of the papers sampled into the “Two–Four” category actually belonged in the “Five” and “Six” categories.

Table A3: Weights associated with the stratified sample.

Strata	Weight (sampled)	Weight (actual)	Final Weight
Two–Four	0.9912	0.9349	0.9471
Five Factors	0.0065	0.0496	0.0424
Six Factors	0.0009	0.0153	0.0103
Seven Factors	0.0014	0.0002	0.0002

Because there are often multiple responses measured on an experiment, and we perform the analysis of sparsity, heredity, and hierarchy using each of these responses, we must also take these multiple responses into account. We did not use all the responses reported in the 129 papers because many of them were highly correlated. Thus, we omitted some (see Section 4). We display the number of responses associated with the actual categorizations of the sampled papers for the responses we retained in Table A4. That is, after papers migrated from their sampled strata to their actual strata there were, for instance, 85 papers with two to four factors and from those papers, there were 128 responses. To obtain the final set of weights in Table A3, we simply reweight the actual paper weights by the expected number of responses, shown in the last column of Table A4. The final weights are in the last column of Table A3 and are used throughout the paper when estimating population quantities.

Table A4: Responses, correctly categorized, in each stratum.

	Papers	Responses	Responses/Paper
Two–Four	85	128	1.506
Five	37	47	1.270
Six	5	5	1
Seven	2	3	1.5

Appendix B. Analysis Results of Experiments Showing No Lack of Fit

Tables B1 through B8 give the results of the regularities analysis on the 115 experiments that showed no significant lack of fit using $\alpha = 0.05$. With the exception of the stationary point analysis, the conclusions made from the full analysis including the 68 experiments that showed significant lack of fit remain unchanged.

Table B1: Effect sparsity, measured as the proportion of active effect types (No LOF).

Effect Type	p-values (SE)	FDR p-values (SE)
Main Effects	0.56 (0.04)	0.52 (0.05)
Two-Factor Interactions	0.27 (0.04)	0.22 (0.04)
Quadratic Effects	0.51 (0.05)	0.44 (0.05)

Table B2: Proportion of active effect types by k for the sample using unadjusted p-values (No LOF). Note these proportions and standard errors do not use the stratified sampling weights.

k	# Experiments	Main Effects (SE)	Two-Factor Interactions (SE)	Quadratic Terms (SE)
2	21	0.47 (0.08)	0.19 (0.09)	0.43 (0.08)
3	47	0.55 (0.04)	0.24 (0.04)	0.47 (0.04)
4	12	0.71 (0.07)	0.29 (0.05)	0.56 (0.07)
5	29	0.58 (0.04)	0.27 (0.03)	0.41 (0.04)
6	4	0.75 (0.09)	0.35 (0.06)	0.54 (0.10)
7	2	0.64 (0.13)	0.36 (0.07)	0.36 (0.13)

Table B3: Proportion of active effect types by k for the sample FDR-adjusted p-values (No LOF). Note these proportions and standard errors do not use the stratified sampling weights.

k	# Experiments	Main Effects (SE)	Two-Factor Interactions (SE)	Quadratic Terms (SE)
2	21	0.36 (0.07)	0.19 (0.09)	0.33 (0.07)
3	47	0.49 (0.04)	0.19 (0.03)	0.39 (0.04)
4	12	0.71 (0.07)	0.21 (0.05)	0.54 (0.07)
5	29	0.43 (0.04)	0.20 (0.02)	0.30 (0.04)
6	4	0.71 (0.09)	0.33 (0.06)	0.54 (0.10)
7	2	0.57 (0.13)	0.33 (0.07)	0.29 (0.12)

Table B4: Estimated average and median effect sizes for the different effect types, weighted to account for stratified sample (No LOF).

Effect Type	Mean (SE)	Median (SE)
Main Effects	5.87 (0.55)	3.54 (0.52)
Two-Factor Interactions	2.00 (0.18)	1.33 (0.14)
Quadratic Effects	4.43 (0.39)	2.50 (0.36)

Table B5: Traditional heredity, measured as the proportion of active two-factor interactions when 0, 1, or 2 parent main effects are active (No LOF).

Strength	p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong	0.43 (0.05)	0.37 (0.06)	$A, B,$ and AB
Weak	0.25 (0.06)	0.21 (0.06)	A and AB
None	0.06 (0.04)	0.05 (0.03)	AB

Table B6: Main effect quadratic heredity, measured as the proportion of active quadratic effects when 0 or 1 of its parent main effects are active (No LOF).

Heredity	p-values (SE)	FDR p-values (SE)	Example of Active Terms
Quadratic Heredity	0.59 (0.06)	0.55 (0.06)	A and A^2
No Quadratic Heredity	0.39 (0.07)	0.31 (0.07)	A^2

Table B7: Quadratic-interaction heredity, measured as the proportion of active two-factor interactions when 0, 1, or 2 of its associated quadratic terms are active (No LOF).

Strength	p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong	0.43 (0.06)	0.40 (0.07)	$A^2, B^2,$ and AB
Weak	0.25 (0.06)	0.20 (0.08)	A^2 and AB
None	0.12 (0.05)	0.09 (0.04)	AB

Table B8: Expanded heredity, measured as the proportion of active interactions with some combination of active main effects and quadratic effects (No LOF).

Strength	p-values (SE)	FDR p-values (SE)	Example of Active Terms
Strong-Strong	0.56 (0.06)	0.50 (0.09)	$A, B, A^2, B^2,$ and AB
Strong-Weak	0.35 (0.08)	0.33 (0.11)	$A, B, A^2,$ and AB
Strong-None	0.25 (0.09)	0.21 (0.08)	$A, B,$ and AB
Weak-Strong	0.38 (0.09)	0.36 (0.11)	$A, A^2, B^2,$ and AB
Weak-Weak	0.23 (0.09)	0.11 (0.07)	$A, A^2,$ and AB or $A, B^2,$ and AB
Weak-None	0.11 (0.07)	0.12 (0.07)	A and AB
None-Strong	0.12 (0.11)	0.19 (0.13)	$A^2, B^2,$ and AB
None-Weak	0.02 (0.02)	0.00 (0.00)	A^2 and AB
None-None	0.04 (0.04)	0.02 (0.02)	AB

[1] Theodore T Allen and Liyang Yu. Low-cost response surface methods from simulation optimization. *Quality and Reliability Engineering International*, 18(1):5–17, 2002.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300, 1995. URL http://www.jstor.org/stable/2346101.

[3] George EP Box and Donald W Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, 1960.

[4] George EP Box and K B Wilson. An analysis for unreplicated fractional factorials. *Journal of the Royal Statistical Society, Series B*, 13(1):1–45, 1951.

[5] Jessica L Chapman, Lu Lu, and Christine M Anderson-Cook. Incorporating response variability and estimation uncertainty into pareto front optimization. *Computers & Industrial Engineering*, 76:253–267, 2014.

[6] Shao-Wei Cheng and CFJ Wu. Factor screening and response surface exploration. *Statistica Sinica*, pages 553–580, 2001.

[7] Hugh Chipman, Michael Hamada, and CFJ Wu. A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, 39(4):372–381, 1997.

[8] Shane Dougherty, James R Simpson, Raymond R Hill, Joseph J Pignatiello, and Edward D White. Effect of heredity and sparsity on second-order screening design performance. *Quality and Reliability Engineering International*, 31(3):355–368, 2015.

[9] Danel Draguljić, David C Woods, Angela M Dean, Susan M Lewis, and Anna-Jane E Vine. Screening strategies in the presence of interactions. *Technometrics*, 56(1):1–1, 2014.

[10] David J Edwards and Robert W Mee. Fractional box-behnken designs for one-step response surface methodology. *Journal of Quality Technology*, 43(4):288, 2011.

[11] David J Edwards and David HQ Truong. A comparison of designs for one-step screening and response surface estimation. *Quality and Reliability Engineering International*, 27(8):1009–1024, 2011.

[12] RL Garrote, RA Bertone, ER Silva, VR Coutaz, and A Avalle. Heat and naoh penetration during chemical peeling of potatoes penetración de calor e hidróxido de sodio durante el pelado químico de patatas. *Food science and technology international*, 4(1):23–32, 1998.

[13] Bradley Jones and Christopher J Nachtsheim. A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, 43(1):1, 2011.

[14] John Lawson. One-step screening and process optimization experiments. *The American Statistician*, 57 (1):15–20, 2003.

[15] Russell V. Lenth. Response-surface methods in R, using rsm. *Journal of Statistical Software*, 32(7):1–17, 2009. URL http://www.jstatsoft.org/v32/i07/.

[16] Xiang Li, Nandan Sudarsanam, and Daniel D Frey. Regularities in data from factorial experiments. *Complexity*, 11(5):32–45, 2006.

[17] Thomas Lumley. survey: Analysis of complex survey samples. http://r-survey.r-forge.r-project.org/survey/, 2016. R package version 3.30.

[18] Christopher J Marley and David C Woods. A comparison of design and model selection methods for supersaturated experiments. *Computational Statistics & Data Analysis*, 54(12):3158–3167, 2010.

[19] William R McDaniel and Bruce E Ankenman. A response surface test bed. *Quality and Reliability Engineering International*, 16(5):363–372, 2000.

[20] Douglas C Montgomery, Raymond H Myers, Walter H Carter, and G Geoffrey Vining. The hierarchy principle in designed industrial experiments. *Quality and reliability engineering international*, 21(2): 197–201, 2005.

[21] Raymond H Myers, Douglas C Montgomery, and Christine M Anderson-Cook. *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons, 2016.

[22] JA Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

[23] John A Nelder. The statistics of linear models: back to basics. *Statistics and Computing*, 4(4):221–234, 1994.

[24] Web of Science. *Web of Science*, 2015. URL http://www.webofknowledge.com.

[25] Maria L Weese, Byran J Smucker, and David J Edwards. Searching for powerful supersaturated designs. *Journal of Quality Technology*, 47(1):66–84, 2015.

[26] CF Jeff Wu and Michael S Hamada. *Experiments: planning, analysis, and optimization*. John Wiley & Sons, 2009.

[27] Ming Yuan, V Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, pages 1738–1757, 2009.