Analysis of Orthogonal Saturated Designs

Daniel T. Voss and Weizhen Wang *

This chapter provides a review of special methods for analyzing data from screening experiments conducted using regular fractional factorial designs. The methods considered are robust to the presence of multiple nonzero effects. Of special interest are methods that try to adapt effectively to the unknown number of nonzero effects. Emphasis is on the development of adaptive methods of analysis of orthogonal saturated designs which rigorously control Type I error rates of tests or confidence levels of confidence intervals under standard linear model assumptions. The robust, adaptive method of Lenth (1989) is used to illustrate the basic problem. Then nonadaptive and adaptive robust methods of testing and confidence interval estimation known to control error rates are introduced and illustrated. While the focus is on Type I error rates and orthogonal saturated designs, Type II error rates, nonorthogonal designs and supersaturated designs are also discussed briefly.

1 Introduction

In the design and analysis of experiments in industry, screening plays an important role in the early phases of experimentation. In Chapter **??**, Montgomery and Jennings provide an overview of screening experiments and also give an introduction to regular fractional factorial designs which are often used in this context. Building upon this foundation, we give further consideration to the analysis of data collected using such designs. In particular, we describe methods that are appropriate when the design of the experiment produces just enough observations to allow estimation of the main effects and interactions of interest; that is, the design is *saturated*. We concentrate on

* Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435, U.S.A. dvoss@math.wright.edu wwang@math.wright.edu

methods of analysis that are *adaptive* in the sense that the estimator of the error variance is altered depending on the values of the estimated main effects and interactions from the experiment.

For motivation and illustration, we consider the plasma etching experiment discussed by Montgomery and Jennings in Chapter **??**. The experimental design was a 2^{6-2}_{IV} regular fractional factorial design with 16 observations that allows the estimation of 15 factorial effects. The data and aliasing scheme are given in Tables 3 and 4 of Chapter **??**.

The linear model, in matrix form, that we use for data analysis is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{1}$$

where the vector $\boldsymbol{Y}$ holds the response variables $Y_1, \ldots, Y_n$; the vector $\boldsymbol{\epsilon}$ holds the error variables $\epsilon_1, \ldots, \epsilon_n$ and these are independent and normally distributed with constant variance σ^2; the vector $\boldsymbol{\beta}$ holds the unknown parameters $\beta_0, \beta_1, \ldots, \beta_h$; and $\boldsymbol{X}$ is the model matrix which is formulated as described below.

The least squares estimate of each main effect is the average of the eight observations at the high level of the factor minus the average of the eight observations at the low level and, likewise, the least squares estimate of each interaction effect is the average of an appropriate set of eight observations minus the average of the other eight observations, as described in Chapter **??**, Section 2. In order for the parameters in our model to measure the main effects and interactions directly, the columns of the model matrix $\boldsymbol{X}$ are formed as follows. The first column consists of a column of ones corresponding to the intercept parameter β_0. For each main effect parameter, the elements of the corresponding column of $\boldsymbol{X}$ consists of $+0.5$ for the high level of the factor and -0.5 for the low level. For the parameter measuring the interaction between factors i and j, the entries in the corresponding column of $\boldsymbol{X}$ are obtained by multiplying the elements of the i and j main effects columns, then multiplying by 2 so that all elements are again $+0.5$ or -0.5.

When an experiment is carefully conducted and the correct model is used, then independence of the response variables is often a reasonable assumption. The experimenters of the plasma etching investigation were reportedly comfortable with the assumptions of normality and constant variance, based on their prior experience with similar experiments. From the 16 observations collected using this design, under the above model assumptions, there are 15 independent factorial effect estimators which can be used for the analysis. We denote these estimators by $\hat{\beta}_i$, $i = 1, 2, \ldots, 15$, corresponding to the effects A, B, C, D, E, F, AB, AD, AE, AF, BD, BE, BF, ABD, ABF. These values of i include a representative effect from each set of aliases for this design; see Chapter **??** for a discussion of aliases and see Table 4 of that chapter for the defining relation and aliases for the plasma etching experiment under consideration here.

Under the assumptions of model (1), the least squares estimators of the 15 main effects and interactions are independently normally distributed with

Table 1. Factorial effect least squares estimates and squared estimates for the plasma etching experiment of Chapter **??**

Effect	$\hat{\beta}_i$	$\hat{\beta}_i^2$
A	-175.50	30800.25
AB	106.75	11395.56
E	103.50	10712.25
B	58.00	3364.00
BE	-53.75	2889.06
ABF	-29.75	885.06
AE	27.25	742.56
D	18.75	351.56
F	-18.75	351.56
C	-18.50	342.25
BF	-16.00	256.00
AF	-13.00	169.00
ABD	-5.75	33.06
AD	4.50	20.25
BD	3.00	9.00

equal variances, and each estimator provides an unbiased estimate of the corresponding effect. The "corresponding effect" is a factorial treatment contrast together with its aliases (as explained in Chapter **??**). Independence of the estimators of these effects is a consequence of using an *orthogonal design*, since an orthogonal design is one that yields uncorrelated estimators under the assumptions of model (1) and uncorrelated estimators are independent under normality. For the plasma etching experiment, the least squares estimates are given in Table 1 of this chapter and will be used to illustrate various methods of analysis.

Given a regular fraction of a 2^f experiment and independent response variables, the estimators described above have constant variance even if the individual response variables do not. Also, the estimators are approximately normally distributed by the Central Limit Theorem. However, if the response variables have unequal variances, this unfortunately causes the estimators to be correlated and, therefore, dependent. The use of the data analysis to assess whether the levels of some factors affect the response variability is, itself, a problem of great interest due to its role in robust product design; see Chapter **??** for a discussion of available methods and analysis. In the present chapter, we consider situations in which the estimators are independent.

One additional premise, fundamental to the analysis of data from screening experiments, is the assumption of *effect sparsity*, namely, that very few of the effects under study are sizable. In a screening experiment, it is common for an investigator to study as many factors as possible in the experiment, but there is usually a restriction on the number of observations that can be run. As a result, screening experiments often include no replication and so provide

no pure estimate of error variance. Furthermore, such experiments are often designed to be saturated (having just enough observations to estimate all of the effects, but leaving no degrees of freedom for error). Thus, there is no mean squared error with which to estimate the error variance independently of effect estimates. The lack of an independent estimator of variance means that standard methods of analysis, such as the analysis of variance and confidence intervals and tests based on the t-distribution, do not apply. Nonetheless, provided that effect sparsity holds, the use of a saturated design often leads to the estimates of the large effects standing out relative to the others; this is fundamental to the effective analysis of data from a saturated design.

A traditional approach to the analysis of data from an orthogonal saturated design utilizes half-normal plots of the effect estimates. This approach was introduced by Daniel (1959) and is illustrated in Figure 3 of Chapter **??** for the plasma etching data. In a half-normal plot, the ordered absolute values of the estimates are plotted against their expected values, or half-normal scores, under the assumption that the estimators all have mean zero in addition to being normally distributed with equal variances. If only a few estimates are relatively large in absolute value, then the corresponding points tend to stand out in the plot away from a line through the points corresponding to the absolute values of the smaller estimates. Daniel advocated using this approach *iteratively*: if the largest estimate is determined, often subjectively, to correspond to a nonzero effect, then that estimate is removed before the next step of the analysis; at the next step, the half-normal plot is regenerated using the remaining effects; the largest remaining estimate is then evaluated. This process is iterated until the largest remaining estimate does not stand out. In practice, such iteration is seldom done. Without iteration, a reasonable interpretation of the half normal plot in Figure 3 of Chapter **??** is that five effects stand out, these being the effects A, AB, E, B, and BE, or their aliases (for example, $BE = AC$).

From an historical perspective, the analysis of orthogonal saturated designs was considered initially by Birnbaum (1959) and Daniel (1959). In addition to half-normal plots for the subjective analysis of orthogonal saturated designs, Daniel (1959) also considered more formal, objective methods of analysis of such designs, as did Birnbaum (1959) in a companion paper. Each considered testing for a nonzero effect amongst h effects, assuming that at most one effect is nonzero. Birnbaum provided a most powerful test of this simple hypothesis (see Section 2 for the definition of a most powerful test). His test is based on the proportion of the total variation that is explained by the largest estimated effect. Such a test could be iterated to test for more than one nonzero effect, but then true error rates and the choice of a best test become unclear. Birnbaum also sought optimal rules for deciding which, and how many, effects are nonzero when at most two effects are truely nonzero and noted that the problem was then already quite complex.

Subsequently, Zahn (1969, 1975ab) considered some variations on the iterative methods of Daniel (1959) and Birnbaum (1959), but his results were

primarily empirical. The subjective use of half-normal plots remains a standard methodology for the analysis of orthogonal saturated designs, but the development of objective methods is progressing rapidly.

Box and Meyer (1986, 1993) provided Bayesian methods for obtaining posterior probabilities that effects are *active*; see Chapter **??**, Section 2, for more details. There followed a flurry of papers proposing new frequentist methods, giving refinements of the methods, and making empirical comparisons of the many variations. Hamada and Balakrishnan (1998) provided an extensive review of these methods, including a Monte Carlo-based comparison of the "operating characteristics" of the methods; that is, a comparison of the power of the methods for a variety of combinations of effect values (*parameter configurations*). They found that comparison of methods is difficult for various reasons. For example, some are intended for individual inferences and others for simultaneous inference. Each method is designed to be "robust", as discussed in Section 3, but each method has its own inherent "breakdown point" with respect to the number of non-negligible effects that can be identified. Still, there are commonalities of the better methods. The preferred methods typically use the smaller ordered absolute values of the estimated effects or the corresponding sums of squares to obtain a robust estimate of variability. We refer the reader to Hamada and Balakrishnan (1998) for further details on the many proposed methods.

The most influential of these methods is a "quick and easy" method introduced by Lenth (1989). Lenth's method was ground breaking because, in addition to being simple, it is *robust* to the presence of more than one large effect and it is *adaptive* to the number of large effects. Furthermore, empirical studies suggest that it maintains good power over a variety of parameter configurations. We use Lenth's method to illustrate the concepts of "robust" and "adaptive" in Section 3 and apply it to the plasma etching study of Chapter **??**. In Section 4, we introduce robust methods of confidence interval estimation, some adaptive and some not, but all known to provide at least the specified confidence level under all parameter configurations. Section 5 contains some analogous results for hypothesis testing. These confidence interval and testing methods are again illustrated using the data from the plasma etching study. The chapter concludes with a broader discussion of the issues for the analysis of unreplicated factorial designs. First though, in Section 2, we give our formulation of the factor screening problem.

2 Formulation of the factor screening problem

For a given experiment with f factors conducted using a 2^{f-q} fractional factorial design, there are $h = 2^{f-q} - 1$ independent factorial effect estimators $\hat{\beta}_i$, $i = 1, \ldots, h$, where $\hat{\beta}_i \sim N(\beta_i, \sigma^2_\beta)$. Suppose that effect sparsity holds so that most of the effects β_i are zero (or negligible), with only a few of the effects being large in magnitude.

The basic factor screening problem is to determine which effects are large and which are small. Any factor found to have only negligible main effects and interactions requires no further investigation and so need not be considered in subsequent experimentation. The primary objective of a screening experiment is, therefore, to screen out unimportant factors so that subsequent experiments can focus on studying the important factors without being unduly large. Also, any factors found to have large effects deserve further study, so their identification is obviously of value.

In the language of hypothesis testing, it is generally agreed that Type I and Type II errors are both of importance in screening experiments. If a Type I error is made, the consequence is that an unimportant factor remains under investigation for one or more subsequent rounds of experimentation, using additional resources and perhaps slowing progress. On the other hand, if a Type II error is made, an important factor may be excluded from future experiments and this could undermine the success of the entire study. In view of this, one could argue that confidence intervals are more useful than hypothesis testing, provided that the confidence intervals are tight enough to pin down adequately the magnitude of each effect. Perhaps rejecting, or failing to reject, the null hypothesis that an effect is zero in the absence of power calculations may be of little value. Or, one might argue that a completely different formulation of the problem is needed, perhaps along the lines of bioequivalence testing (see Brown, Hwang, and Munk, 1997), where the goal is to demonstrate similarity of effects rather than differences. For example, one could try to demonstrate that certain effects are close to zero so that the corresponding factors merit no further consideration.

The above discussion suggests possible difficulties in formulating the problem of data analysis for screening experiments. However, it is perhaps more accurate to say that, even if Type II errors are as, or more important than, Type I errors in the analysis of screening experiments, it is more difficult to deal with Type II errors. (This is because the probability of making Type II errors depends on the parameter configuration under composite alternative hypotheses.) We can obviously avoid making Type II errors by always asserting that all effects are nonzero, so never screening out any factors, but then the primary goal of a screening experiment cannot be achieved. So, in searching for the methods which are best at detecting large effects in the analysis of screening experiments, one must strike a balance between Type I and Type II errors.

To be able to compare methods even-handedly, we have chosen to rely on a fundamental premise in statistical hypothesis testing, namely, we seek the most powerful level-α test. A test is of *level-α* if the probability of making a Type I error is at most α. If this upper bound on the Type I error rate is the least upper bound (as it is when the upper bound is achieved for at least one set of parameter values), the test is said to be of *size* α. Amongst level-α tests, one is a *most powerful test* if, loosely speaking, it is the most likely to detect nonzero effects, large or small. In the analysis of saturated designs,

establishing that a test is level-α is complicated by the fact that all of the effect parameters are of interest yet, in testing any single effect, the other effects are nuisance parameters. For the methods of analysis proposed in the literature, critical values are invariably determined under the assumption that all of the effects β_i are zero; we call this the *null case*. However, for most of the available methods of data analysis, it still remains an open problem to prove that use of critical values obtained under the null distribution yields a level-α test. In other words, for most proposed methods of data analysis, it remains an open problem to show that the null case is the "least favorable"; that is, it is the parameter configuration that has the highest Type I error rate. If this is true, then the use of these critical values will yield size-α tests. Such may well be the case for orthogonal designs but, most curiously, Wang and Voss (2001) provided a counterexample for a particular method of analysis of nonorthogonal saturated designs.

In view of these considerations, a reasonable goal is to seek methods of analysis for screening experiments that include powerful tests of specified size or level and exact confidence intervals that are tight. An *exact confidence interval* is analogous to a test of specified size—"exact" means that the confidence level is at least as high as the level claimed and the level claimed is the greatest lower bound on the confidence level.

3 Robust adaptive methods

The most influential method of analysis of orthogonal saturated designs yet proposed is the robust, adaptive method of Lenth (1989). The "quick and easy" method that he proposed is based on the following estimator of the standard deviation, σ_β, of the effect estimators $\hat{\beta}_i$. This estimator is "robust" and "adaptive", concepts which will be explained in detail after the following description of the method.

First, obtain an initial estimate of σ_β:

$$\hat{\sigma}_o = 1.5 \times (\text{the median of the absolute estimates } |\hat{\beta}_i| \,). \tag{2}$$

For the plasma etching experiment, the median absolute estimate is 18.75, which can be found as the eighth listed value in Table 1. This yields $\hat{\sigma}_o = 28.13$ from (2). If the estimators $\hat{\beta}_i$ are all normally distributed with mean zero and common standard deviation σ_β, then $\hat{\sigma}_o$ is approximately unbiased for σ_β.

Secondly, calculate an updated estimate

$$\hat{\sigma}_L = 1.5 \times (\text{the median of those } |\hat{\beta}_i| \text{ that are less than } 2.5\hat{\sigma}_o),$$

where subscript L denotes Lenth's method. For the plasma etching experiment, $2.5\hat{\sigma}_o = 70.31$. Each of the three largest absolute estimates exceeds this value and so is set aside. The median of the remaining 12 absolute estimates is 18.63. Thus $\hat{\sigma}_L = 27.94$, which is slightly smaller than $\hat{\sigma}_o$.

Lenth (1989) referred to this estimator $\hat{\sigma}_L$ as the *pseudo standard error* of the estimators $\hat{\beta}_i$. He recommended using the test statistic $\hat{\beta}_i/\hat{\sigma}_L$ to test the null hypothesis $\beta_i = 0$ and recommended using the quantity $(\hat{\beta}_i - \beta_i)/\hat{\sigma}_L$ to construct a confidence interval for β_i.

Critical values for individual tests and confidence intervals are based on the *null distribution* of $|\hat{\beta}_i|/\hat{\sigma}_L$; that is, on the distribution of this statistic when all effects β_i are zero. Lenth proposed a t-distribution approximation to the null distribution, whereas Ye and Hamada (2000) obtained exact critical values by simulation of $|\hat{\beta}_i|/\hat{\sigma}_L$ under the null distribution. From their tables of critical values, the upper 0.05 quantile of the null distribution of $|\hat{\beta}_i|/\hat{\sigma}_L$ is $c_L = 2.156$. On applying Lenth's method for the plasma etching experiment and using $\alpha = 0.05$ for individual inferences, the minimum significant difference for each estimate is calculated to be $c_L \times \hat{\sigma}_L = 60.24$. Hence, the effects A, AB and E are declared to be nonzero, based on individual 95% confidence intervals.

From empirical comparisons of various proposed methods of analysis of orthogonal saturated designs, (Hamada and Balakrishnan, 1998; Wang and Voss, 2003), Lenth's method can be shown to have competitive power over a variety of parameter configurations. It remains an open problem to prove that the null case is the least favorable parameter configuration.

We now discuss what it means for a method of analysis to be "robust" or "adaptive". Lenth's method is *adaptive* because of the two-stage procedure used to obtain the pseudo standard error. The pseudo standard error $\hat{\sigma}_L$ is computed from the median of most of the absolute estimates, but how many are excluded from the calculation depends on the data. In other words, the procedure adapts itself to the data, and it attempts to do so efficiently. It seems reasonable to believe that $\beta_i \neq 0$ if $|\hat{\beta}_i| > 2.5\hat{\sigma}_o$ because, for a random variable X having a normal distribution with mean zero and standard deviation σ, $P(|X| > 2.5\sigma) \approx 0.0124$. In a sense, one is pre-testing each hypothesis

$$H_{0,i} : \beta_i = 0$$

in order to set large estimates aside in obtaining the pseudo standard error, which is then used for inference on the remaining effects β_i.

Consider now robustness. If the estimators $\hat{\beta}_i$ are computed from independent response variables then, as noted in Section 1, the estimators have equal variances and are usually at least approximately normal. Thus the usual assumptions, that estimators are normally distributed with equal variances, are approximately valid and we say that there is inherent *robustness* to these assumptions. However, the notion of *robust methods* of analysis for orthogonal saturated designs refers to something more. When making inferences about any effect β_i, all of the other effects β_k $(k \neq i)$ are regarded as nuisance parameters and "robust" means that the inference procedures work well, even when several of the effects β_k are large in absolute value. Lenth's method is robust because the pseudo standard error is based on the median absolute estimate and hence is not affected by a few large absolute effect estimates.

The method would still be robust even if one used the initial estimate $\hat{\sigma}_o$ of σ_β, rather than the adaptive estimator $\hat{\sigma}_L$, for the same reason.

Any robust method has a *breakdown point*, which is the percentage of large effects which would make the method ineffective. For Lenth's method, if half or more of the effect estimates are very large in magnitude, then $\hat{\sigma}_o$ will be large and hence so will $\hat{\sigma}_L$, causing the method to lose so much power that the method breaks down. Hence, the breakdown point is about 50%. One could lower the breakdown point by using, for example, the 30th percentile of the absolute estimates rather than the median to estimate σ_β. However, this would increase the variability of the pseudo standard error, which would reduce power when there truly is effect sparsity.

In summary, Lenth's method is robust in the sense that it maintains good power as long as there is effect sparsity and it is adaptive to the degree of effect sparsity, using a pseudo standard error that attempts to involve only the estimates of negligible effects.

Like many methods of analysis of orthogonal saturated designs proposed in the literature, the critical values for Lenth's method are obtained in the null case (all β_i zero), assuming this is sufficient to control the Type I error rates. This raises the question: can one establish *analytically* that Lenth's and other proposed methods do indeed provide the claimed level of confidence or significance under standard model assumptions? The rest of this chapter concerns methods for which the answer is "yes".

4 Robust exact confidence intervals

In this section, we discuss the construction of individual confidence intervals for each factorial effect β_i, based only on the least squares estimates $\hat{\beta}_1, \ldots, \hat{\beta}_h$. An exact $100(1-\alpha)\%$ confidence interval for β_i is analogous to a size-α test of the null hypothesis $\beta_i = 0$, against a two-sided alternative. In testing this hypothesis, the probability of making a Type I error depends on the values of the other parameters β_k, $k \neq i$. Such a test would be of size α, for example, if under the null hypothesis, the probability of making a Type I error were exactly α when $\beta_k = 0$ for all $k \neq i$ and at most α for any other values of the β_k for $k \neq i$. By analogy, a confidence interval for β_i would be an exact $100(1-\alpha)\%$ confidence interval if the confidence level were exactly $100(1-\alpha)\%$ when $\beta_k = 0$ for all $k \neq i$ and at least $100(1-\alpha)\%$ for any values of the β_k for $k \neq i$. Inference procedures that control the error rates under all possible parameter configurations are said to provide *strong control of error rates*; see Hochberg and Tamhane (1987) and also Chapter **??**. For a confidence interval, the error rate is at most α if the confidence level is at least $100(1-\alpha)\%$.

When screening experiments are used, it is generally anticipated that several effects may be nonzero. Hence, one ought to use statistical procedures that are known to provide strong control of error rates. It is not enough to

control error rates only under the complete null distribution. This section discusses exact confidence intervals. Size-α tests are considered in Section 5.

4.1 Non-adaptive confidence intervals

The first confidence interval for the analysis of orthogonal saturated designs which provided strong control of the error rate was established by Voss (1999). His confidence interval for β_i excludes $\hat{\beta}_i$ from the computation of the standard error and is obtained using the random variable

$$(\hat{\beta}_i - \beta_i)/\hat{\sigma}_V \,,$$

where the denominator is the square root of

$$\hat{\sigma}_V^2 = \frac{\sum_{k=1}^{u} \hat{\beta}_{(k)}^2}{u} \,, \tag{3}$$

which is the mean squared value of the u smallest of the $h-1$ effect estimates excluding $\hat{\beta}_i$, and where u is specified before the data are examined. Here $\hat{\beta}_{(k)}^2$ denotes the kth smallest of the $h-1$ squared estimates $\hat{\beta}_k^2$ for $k \neq i$.

Let c_V be the upper-α critical value, obtained as the upper-α quantile of the distribution of $|\hat{\beta}_i|/\hat{\sigma}_V$ when all effects are zero. Voss (1999) showed that

$$\hat{\beta}_i \pm c_V \hat{\sigma}_V$$

is an exact $100(1-\alpha)\%$ confidence interval for β_i. This result was obtained from a basic but obscure *Stochastic Ordering Lemma*, which says that

$$|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \tag{4}$$

is stochastically largest under the complete null distribution. This follows because (4) is a non-increasing function of $\hat{\beta}_k^2$ for each $k \neq i$, the estimators $\hat{\beta}_k$ $(k = 1, \ldots, h)$ are independent, and the distribution of each $\hat{\beta}_k^2$ $(k \neq i)$ is increasing in β_k^2. As a consequence, if

$$P\left(|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \leq c_V\right) = 1 - \alpha$$

under the null distribution, then

$$P\left(|\hat{\beta}_i - \beta_i|/\hat{\sigma}_V \leq c_V\right) \geq 1 - \alpha$$

under any parameter configuration. This stochastic ordering result was obtained independently by Alam and Rizvi (1966) and Mahamunulu (1967).

We now apply Voss' method to the data from the plasma etching experiment to construct *individual* 95% confidence intervals for each effect using $u = 8$. The pooling of 8 sums of squares into the denominator (3) provides

a reasonably robust procedure without undue loss of power. One could, of course, pool more than 8 sums of squares into the denominator, as one would usually anticipate greater effect sparsity—more than 8 negligible effects—in a screening experiment. Still, 8 provides a reasonable tradeoff between power and robustness. Also, for simultaneous confidence intervals, Dean and Voss (1999) provided critical values for this choice because, in a single replicate 2^4 factorial experiment, an inactive factor is involved in eight inactive main effects and interactions which could then be used to provide the denominator (3) in Voss' method.

On application of Voss' method for $h = 15$ total effects, $u = 8$ smallest effects and 95% individual confidence level, we find by simulation that the critical value is $c_V = 5.084$. This and subsequent simulated critical values in this chapter are obtained by generating a large number of sets of estimates under the null distribution (that is, with mean zero and standard deviation one), computing the relevant statistic for each set of estimates, and using the upper α quantile of the resulting empirical distribution of values of the statistic as the critical value. Sample programs for computing critical values are available at http://www.wright.edu/~dan.voss/.

For any of the seven largest estimates, expression (3) gives $\hat{\sigma}_V^2 = 191.59$, so the minimum significant difference is $c_V \hat{\sigma}_V = 70.37$. From Table 1, we see that three effects have estimates larger than 70.37 in absolute value, namely, A, AB and E. Hence, individual 95% confidence intervals for these three effects do not include zero, so these effects are declared to be non-zero. No other effects can be declared to be non-zero using this method. These results match those obtained using Lenth's individual 95% confidence intervals.

Voss and Wang (1999) showed that *simultaneous* confidence intervals could be obtained using a similar, but more complicated, technical justification. For simultaneous intervals, the computation of the critical value is based on the null distribution of the maximum of the h random variables $|\hat{\beta}_i|/\hat{\sigma}_V$ and the upper α quantiles can be obtained via simulation, as described above. . The critical values are provided in Table A.11 of Dean and Voss (1999). We do not illustrate this method here but instead illustrate adaptive simultaneous confidence intervals in the following section. Based on power simulations conducted by Wang and Voss (2003), adaptive methods appear to have better minimax power and competitive average power when compared to non-adaptive methods over a variety of parameter configurations.

4.2 Adaptive confidence intervals

We now extend the above ideas to obtain adaptive individual confidence intervals for each effect β_i and again apply the methods to the plasma etching example. In developing such intervals that strongly control the error rate, the motivation of Wang and Voss (2001, 2003) was the approach used when examining the half-normal probability plot. This involves looking for a jump in the magnitude of the absolute estimates, or their squared values, in order to

determine how many of these should be pooled into the estimate of σ_β^2. Below, we describe the methodology of Wang and Voss (2003) and, for simplicity, we concentrate on the special case of their general theory that is most useful in practice.

Suppose that one allows the possibility of pooling into the estimate of error the j smallest of the $h-1$ squared estimates, excluding $\hat{\beta}_i^2$, for some prespecified set J of choices for j. For example, for $h = 15$ effects, one might consider pooling either 8 or 12 of the 14 available squared estimates, since 12 might give very good power if only one or two effects are non-negligible, but 8 would be a better choice if there happens to be less effect sparsity. This corresponds to taking $J = \{8, 12\}$, and simulations by Wang and Voss (2003) found that this choice provides good power under a variety of parameter configurations. Let

$$\hat{\sigma}_j^2 = w_j \sum_{k=1}^{j} \hat{\beta}_{(k)}^2 / j \tag{5}$$

denote the mean squared value of the j smallest squared estimates (excluding $\hat{\beta}_i$), scaled by a prespecified weight w_j, and let $\hat{\sigma}_{\min}^2$ be the minimum value of all the $\hat{\sigma}_j^2$ for values of j in the prechosen set J; that is,

$$\hat{\sigma}_{\min}^2 = \min\{\hat{\sigma}_j^2 \mid j \in J\}\,. \tag{6}$$

Since the $\hat{\beta}_{(k)}^2$ are ordered in increasing value, $\sum_{k=1}^{j} \hat{\beta}_{(k)}^2/j$ is increasing in j. So, the condition that $w_j < w_{j'}$ for $j > j'$ is imposed on the constants w_j in order for $\hat{\sigma}_{\min}$ to be adaptive, namely, so that any j could yield the minimum value of $\hat{\sigma}_j^2$. Also, $\hat{\sigma}_{\min}$ is a non-decreasing function of each $\hat{\beta}_k^2$ for $k \neq i$, providing the means to establish strong control of error rates via the Stochastic Ordering Lemma.

An application of this lemma shows that an exact $100(1-\alpha)\%$ confidence interval for β_i is given by

$$\hat{\beta}_i \pm c_{\min}\hat{\sigma}_{\min}\,,$$

where $c_{\min}$ denotes the upper-α quantile of the null distribution of $|\hat{\beta}_i|/\hat{\sigma}_{\min}$.

Some further guidance is needed concerning specification of the set J and the weights w_j for $j \in J$. When exactly j of the effects are zero or negligible, it is desirable that $\hat{\sigma}_{\min}$ is equal to $\hat{\sigma}_j$ and the chance of this happening is greater for smaller w_j. This provides some basis for choosing the w_j using any existing knowledge concerning the likely number of negligible effects. However, one generally does not know how many effects are negligible; hence the desire for a robust, adaptive method. Wang and Voss (2003) conducted an empirical power study of the procedure for various choices of J and w_j for $j \in J$ for the analysis of 15 effects. A choice of J that yielded good minimax and average power over a variety of parameter configurations was the set $J = \{8, 12\}$, for which either the 8 or 12 smallest squared estimates are pooled to estimate

the variance. Furthermore, for this choice, each weight w_j, for $j \in \{8, 12\}$, was chosen to make $\hat{\sigma}_j^2$ an unbiased estimator of the common variance of the estimators $\hat{\beta}_i$ when all effects β_i are zero. To apply this method, each value w_j can be obtained by simulation by computing the average value of $\sum_{k=1}^{j} Z_{(k)}^2/j$—that is, the average of the j smallest squared values of $h-1$ pseudo standard normal random variables—and then taking the reciprocal of this average.

If we apply this method to the plasma etching experiment and compute *individual* 95% confidence intervals using $J = \{8, 12\}$, we obtain by simulation the values $w_8 = 4.308$, $w_{12} = 1.714$ and $c_{\min} = 2.505$. For the effects A, AB and E corresponding to the three largest estimates, $\hat{\sigma}_8^2 = 825.35$ and $\hat{\sigma}_{12}^2 = 1344.54$, so $\hat{\sigma}_{\min}^2 = 825.35$ and the minimum significant difference is $c_{\min}\hat{\sigma}_{\min} = 71.97$. Therefore the effects A, AB and E are declared to be significantly different from zero. The effects with the next largest estimates are B and AC. These are not significant based on the same minimum significant difference, although the value of $\hat{\sigma}_{12}^2$ (used to compute the corresponding test statistics **NOT TRUE???**) is larger for assessing these effects, since it is computed using the 12 smallest squared estimates apart from the one for which the confidence interval is being constructed.

Wang and Voss (2003) showed that *simultaneous* confidence intervals could be obtained in a similar way, by computing the critical value based on the null distribution of the maximum of the h random variables $|\hat{\beta}_i|/\hat{\sigma}_{\min}$, $i = 1, \ldots, h$, where, for each i, $\hat{\sigma}_{\min}$ is a function of the estimators excluding $\hat{\beta}_i$. To obtain simultaneous 95% confidence intervals for all 15 effects, the simulated critical value provided by Wang and Voss (2003) is $c_{\min} = 6.164$. For examining each of the 7 largest estimates, $\hat{\sigma}_{\min}^2$ is again equal to $\hat{\sigma}_8^2 = 825.35$ from (5) and (6). So we find that the minimum significant difference for simultaneous 95% confidence intervals is

$$c_{\min}\hat{\sigma}_{\min} = 177.08.$$

The largest effect estimate, $\hat{\beta}_A = -175.50$, has magnitude just under this more stringent threshold for simultaneous inference and thus none of the effects are found to differ significantly from zero.

Obviously, simultaneous 95% confidence intervals are more conservative than individual 95% confidence intervals, explaining the lack of significant results in this case. We advocate that both individual and simultaneous confidence intervals be used, since they provide different information and are both useful. The finding that three effects are significant using individual, but not simultaneous, 95% confidence intervals suggests the possibility of false positives, for example, whereas the significant results would be more believable if the simultaneous confidence intervals identified the same effects as being significant.

Simulations of Wang and Voss (2003) show little difference in power between their adaptive confidence intervals using $J = \{8, 12\}$ and the confidence

intervals of Lenth (1989), though the former have the advantage that control of Type I error rates is established.

5 Robust size-α tests

In this section we focus on hypothesis testing and discuss both individual and simultaneous tests for detecting non-zero effects. Of special interest are the step-down tests described in Section 5.2, as these offer improved power over single-step tests.

5.1 Individual and simultaneous single-step tests

Adaptive, robust *single-step tests* of size α, both individual and simultaneous, can be based on the corresponding confidence intervals already discussed. To test the hypothesis

$$H_{0,i} : \beta_i = 0$$

for each fixed i, or to test these hypotheses simultaneously, one may simply check whether the corresponding Wang and Voss (2003) individual or simultaneous confidence intervals include zero. The test procedure is: *reject each null hypothesis $H_{0,i}$ if and only if the confidence interval for β_i excludes zero.* This testing procedure controls the error rate and uses the data adaptively.

Better yet, one can obtain adaptive, robust tests that are more easily implemented and are still of the specified size. For testing each null hypothesis $H_{0,i} : \beta_i = 0$, one need not exclude the corresponding estimator $\hat{\beta}_i$ from the computation of the denominator or standard error, since $\beta_i = 0$ under the null hypothesis.

For example, to obtain an individual test of $H_{0,i} : \beta_i = 0$, Berk and Picard (1991) proposed rejecting the null hypotheses for large values of the test statistic

$$\frac{\hat{\beta}_i^2}{\sum_{k=1}^{u} |\hat{\beta}|_{(k)}^2 / u},$$

where the denominator is the mean value of the u smallest squared estimates computed from all h estimates, for a prespecified integer u. This test controls the error rate because the test statistic is non-increasing in $\hat{\beta}_k^2$ for each $k \neq i$. Also, since the denominator is the same for testing each hypothesis $H_{0,i}$, implementation of the test is simple relative to implementation of corresponding confidence interval procedures.

Analogously, Voss and Wang (2005) proposed individual and simultaneous adaptive tests based on $\hat{\beta}_i / \hat{\sigma}_{\min}$ for $i = 1, \ldots, h$, which are similar to the adaptive confidence intervals of Section 4.2, but with $\hat{\sigma}_{\min}$ computed from all h estimators $\hat{\beta}_k$ rather than by setting aside $\hat{\beta}_i$ when testing $H_{0,i} : \beta_i = 0$. The more powerful version of this test will be discussed in Section 5.2.

5.2 Step-down tests

While a single-step test compares each effect estimate with the same critical value, a *step-down test* uses this "single-step" critical value only for the largest effect estimate, then "steps down" to test the next largest effect estimate using a sharper critical value, stepping down iteratively and stopping only when an effect is not significant. It is well known, by virtue of sharper critical values after testing the effect with largest estimate, that simultaneous step-down tests have a clear power advantage over simultaneous single-step tests; see, also, Chapter **??**.

Although step-up tests are analogous to step-down tests, they are not considered here since error rate control remains an open problem. Step-down or step-up tests have been proposed for the analysis of orthogonal saturated designs by Voss (1988), Voss and Wang (2005), Venter and Steel (1996, 1998), Langsrud and Naes (1998), and Al-Shiha and Yang (1999). Here we provide the details of the tests of Voss and Wang (2005) since they have been proved to control the error rate. The other tests are intuitively attractive but have been 'justified' only empirically.

To develop the test, we use the *closed testing procedure* of Marcus, Peritz, and Gabriel (1976). This procedure requires the construction of a size-α test of the hypothesis

$$H_{0,I} : \beta_i = 0, \text{ for all } i \in I$$

for each non-empty index set $I \subset \{1, \ldots, h\}$. We test this null hypothesis using the test statistic

$$T_I = \max_{i \in I} T_i \,, \qquad \text{where } T_i = \hat{\beta}_i^2 / \hat{\sigma}_{\min}^2 \,, \tag{7}$$

and where $\hat{\sigma}_{\min}^2$ is defined as in equation (6) but with the modification that each $\hat{\sigma}_j^2$ is computed using all h effect estimators $\hat{\beta}_i$, rather than setting one aside.

Let c_I denote the upper-α quantile of the distribution of T_I when all h effects are zero. Then, the test that rejects $H_{0,I}$ if $T_I > c_I$ is a size-α test of the null hypothesis since the test statistic T_I is a non-increasing function of $\hat{\beta}_k^2$ for each $k \notin I$ (Voss and Wang, 2005). For each i, the closed testing procedure rejects $H_{0,i} : \beta_i = 0$ if and only if $H_{0,I}$ is rejected for each I containing i. Use of this procedure controls the simultaneous error rate to be at most α; see Marcus, Peritz, and Gabriel (1976). It requires the testing of $2^h - 1$ hypotheses, one for each subset I of effects. It is then necessary to sort through the results to determine which effects β_i can be declared to be non-zero. However, by definition of the test statistic in (7), $I \subset I'$ implies that $T_I \leq T_{I'}$. Also, it can be shown that the critical values c_I decrease as the size of the set I decreases. Thus, we can obtain a "shortcut" as follows, (see also Chapter **??**).

Step-down Test Procedure: Let $[1], \ldots, [h]$ be the indices of the effects after reordering so that $T_{[1]} < \cdots < T_{[h]}$. We denote by $c_{j,\alpha}$ the upper-α critical value c_I for any index set I of size j. The steps of the procedure are:

S1: If $T_{[h]} > c_{h,\alpha}$, then infer $\beta_{[h]} \neq 0$ and go to step 2; else stop.
S2: If $T_{[h-1]} > c_{h-1,\alpha}$, then infer $\beta_{[h-1]} \neq 0$ and go to step 3; else stop.
S3: $\cdots$

This procedure typically stops within a few steps due to effect sparsity. Voss and Wang (2004) proved, for the above test of $H_{0,i} : \beta_i = 0$ $(i = 1, \ldots, h)$, that the probability of making any false inferences (Type I) is at most α for any values of the parameters $\beta_1, \ldots, \beta_h$.

We now apply the step-down test to the plasma etching experiment, choosing a *simultaneous* significance level of $\alpha = 0.05$ and $J = \{8, 12\}$, as described in Section 4.2. We obtained, via simulation, the values $w_8 = 4.995$, $w_{12} = 2.074$, $c_{15,0.05} = 4.005$, $c_{14,0.05} = 3.969$, and subsequent critical values not needed here. The values of w_8 and w_{12} are different from those in Section 4.2 for confidence intervals, since now no $\hat{\beta}_i$ is set aside to obtain $\hat{\sigma}_{\min}$. For testing *each* effect, $\hat{\sigma}_8^2 = 956.97$ and $\hat{\sigma}_{12}^2 = 1619.94$, so that $\hat{\sigma}_{\min}^2 = 956.97$. Hence, the minimum significant difference for the largest estimate is $c_{15,0.05}\hat{\sigma}_{\min} = 123.89$, and A is declared to be significantly different from zero. Stepping down, the minimum significant difference for the second largest estimate is $c_{14,0.05}\hat{\sigma}_{\min} = 122.78$ and AB is not declared to be significantly different from zero, nor are any of the remaining effects at the simultaneous 5% level of significance.

We recommend the use of this method for the analysis of orthogonal saturated designs. It is the only adaptive step-down test in the literature known to control Type I error rates. Furthermore, it is more powerful than the corresponding single-step test. The single-step test is analogous to the simultaneous confidence intervals of Wang and Voss (2003) and the latter were shown via simulation to provide good minimax and average power over a variety of parameter configurations.

6 Discussion

There are important advantages in using adaptive robust procedures that strongly control the error rate. Strong control of the error rate provides the statistical rigor for assessing the believability of any assertions made about the significance of the main effects or interactions, whether confidence intervals or tests are applied. Use of a robust adaptive procedure allows the data to be used efficiently.

From the mathematical viewpoint, the existence of robust adaptive procedures that strongly control the error rates is facilitated by the availability of an adaptive variance estimator which is stochastically smallest when all β_i are zero (the null case). Wang and Voss (2003) provided a large class of such

robust adaptive estimators, with a lot of flexibility in the possible choices of the set J and weights w_j for construction of adaptive variance estimates. A remaining problem is to determine which estimators in the class are the most robust, in the sense of performing well over all reasonable parameter configurations under effect sparsity. Additional simulation studies seem necessary to investigate this issue; see Wang and Voss (2003).

One way to construct or formulate an *adaptive variance estimator* is to have multiple variance estimators and to have a data-dependent choice of which one is used. Adaptive methods known to strongly control error rates use special variance estimators of this type. In particular, the construction of adaptive variance estimators that allow strong control of error rates depends fundamentally on choosing between several possible variance estimators where (i) each possible variance estimator is a non-increasing function of each squared estimator $\hat{\beta}_i^2$, and (ii) the adaptive estimator is the minimum of these contending variance estimators. Under these circumstances, this minimum is also non-increasing in each squared estimator $\hat{\beta}_i^2$, as required for application of the Stochastic Ordering Lemma. The lemma also requires that the estimators $\hat{\beta}_i$ be independent and that the distribution of each squared estimator $\hat{\beta}_i^2$ be non-decreasing in β_i^2. These requirements all hold for analysis of an orthogonal design under standard linear model assumptions. Under such assumptions, an orthogonal design yields independent effect estimators. A design is *nonorthogonal* if any of the estimators $\hat{\beta}_i$ are correlated under standard linear model assumptions.

For a nonorthogonal design, the problem of how to construct a variance estimator that is robust, adaptive and known to strongly control error rates is difficult. So far, there is only one procedure known to strongly control error rates. For analysis of nonorthogonal designs, this method, developed by Kinateder, Voss and Wang (1999), builds upon a variance estimation approach of Kunert (1997) using sequential sums of squares. However, the use of sequential sums of squares is equivalent to making a linear transformation of the estimators $\hat{\beta}_i$ to obtain independent estimators $\hat{\tau}_i$, for which the corresponding sequential sums of squares are also independent. Unfortunately, if the effects are entered into the model in the order $\beta_1, \beta_2, \ldots$, then the expected value of $\hat{\tau}_i$ can involve not only τ_i but also the effects τ_j for $j \geq i$; see Kunert (1997). As a consequence, effect sparsity is diminished for the means of the transformed estimators, $\hat{\tau}_i$.

The problem of data analysis is even harder for supersaturated designs. Then, not only is there necessarily nonorthogonality, but estimability of effects also becomes an issue. Chapter **??** discusses some of the serious problems involved in the analysis of supersaturated designs. Related references include Abraham, Chipman, and Vijayan (1999), Lin (2000), and Holcomb, Montgomery, and Carlyle (2003). There is currently no method of analysis of supersaturated designs that is known to provide strong control of error rates. Further work is needed in this area, as well as on the analysis of nonorthogonal saturated designs.

Returning to the discussion of orthogonal saturated designs, Lenth's (1989) method works well in general and no simulation study so far has detected a parameter configuration for which the error rate is not controlled. Also, his procedure is not very dependent on making a good initial guess for the number of negligible effects, although use of the median causes his method to break down if more than half the effects are large. It is of great interest to show that his method strongly controls the error rate. We cannot use the adaptive technique developed by Wang and Voss (2003) to resolve this question for Lenth's method because their method uses a monotone function of the absolute effect estimates to estimate σ_β, whereas Lenth's method and its variants, proposed by Dong (1993) and Haaland and O'Connell (1995), do not.

In the search for non-negligible effects under effect sparsity, it may seem more reasonable to step up than to step down; that is, to start with evaluation of the smaller estimates, and step up from the bottom until the first significant jump in the estimates is found. Then all effects corresponding to the large estimates could be declared non-zero. Step-down tests, where the largest estimates are considered first and one steps down as long as each estimate in turn is significant, are often justifiable by the closure method of Marcus, Peritz, and Gabriel (1976). However, the mathematical justification of step-up tests, including those of Venter and Steel (1998) and others, remains an interesting and open issue. Wu and Wang (2004) have made some progress in this direction and have provided a step-up test for the number of nonzero effects that provides strong control of error rates. If at least three effects are found to be nonzero, one would like to conclude that the three effects corresponding to the three largest estimates are nonzero, but the procedure does not provide this guarantee. Clearly further research on step-up procedures is needed.

It is appropriate to close this chapter with the following reminder. Although the focus of the work described here is on controlling the Type I error rate for adaptive robust methods, the problem of Type II errors in the analysis of screening experiments should not be overlooked. If a Type II error is made, then an important factor would be ignored in subsequent experiments. On the other hand, if a Type I error is made, then an inactive factor is unnecessarily kept under consideration in future experiments, which is less serious. We argue that the best tests are the most powerful tests of specified size. In a simulation study, we showed (Wang and Voss, 2003) that, in terms of power, adaptive methods known to have strong control of Type I error rates are competitive with alternative methods for which strong control of Type I error rates remains to be established. Hence, one can use adaptive robust methods known to strongly control error rates without sacrificing power.

Acknowledgements

The authors are grateful to anonymous reviewers for their helpful comments. This research was supported by National Science Foundation Grant No. DMS-0308861.

References

Abraham, B., Chipman, H. and Vijayan, K. (1999). Some risks in the construction and analysis of supersaturated designs. *Technometrics*, **41**, 135-141.

Alam, K. and Rizvi, M. H. (1966). Selection from multivariate normal populations. *Annals of the Institute of Statistical Mathematics*, **18**, 307–318.

Al-Shiha, A. A. and Yang, S.-S. (1999). A multistage procedure for analyzing unreplicated factorial experiments. *Biometrical Journal*, **41**, 659–670.

Berk, K. N. and Picard, R. R. (1991). Significance tests for saturated orthogonal arrays. *Journal of Quality Technology*, **23**, 79–89.

Birnbaum, A. (1959). On the analysis of factorial experiments without replication. *Technometrics*, **1**, 343–357.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Box, G. E. P. and R. D. Meyer (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology*, **25**, 94–105.

Brown, L. D., Hwang, J. T. G. and Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, **26**, 2345–2367.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311–341.

Dean, A. M. and Voss, D. T. (1999). *Design and Analysis of Experiments*. Springer, New York.

Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, **3**, 209–217.

Haaland, P. D. and O'Connell, M. A. (1995). Inference for effect-saturated fractional factorials. *Technometrics*, **37**, 82–93.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, **8**, 1–41.

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

Holcomb, D. R., Montgomery, D. C. and Carlyle, W. M. (2003). Analysis of supersaturated designs. *Journal of Quality Technology*, **35**, 13-27.

Kinateder, K. J., Voss, D. T. and Wang, W. (1999). Exact confidence intervals in the analysis of nonorthogonal saturated designs. *American Journal of Mathematical and Management Sciences*, **20**, 71–84.

Kunert, J. (1997). On the use of the factor-sparsity assumption to get an estimate of the variance in saturated designs. *Technometrics*, **39**, 81–90.

Langsrud, O. and Naes, T. (1998). A unified framework for significance testing in fractional factorials. *Computational statistics and data analysis*, **28**, 413–431.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Lin, D. K. J. (2000). Recent developments in supersaturated designs. In *Statistical Process Monitoring and Optimization*, Chapter 18. Editors: S. H. Park and G. G. Vining. Marcel Dekker, New York, 305-319.

Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. *Annals of Mathematical Statistics*, **38**, 1079–1091.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Venter, J. H. and Steel, S. J. (1996). A hypothesis-testing approach toward identifying active contrasts. *Technometrics*, **38**, 161–169.

Venter, J. H. and Steel, S. J. (1998). Identifying active contrasts by stepwise testing. *Technometrics*, **40**, 304–313.

Voss, D. T. (1988). Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error. *Communications in Statistics: Theory and Methods*, **17**, 3345–3359.

Voss, D. T. (1999). Analysis of orthogonal saturated designs. *Journal of Statistical Planning and Inference*, **78**, 111-130.

Voss, D. T. and Wang, W. (1999). Simultaneous confidence intervals in the analysis of orthogonal saturated designs. *Journal of Statistical Planning and Inference* **81**, 383–392.

Voss, D. T. and Wang, W. (2005). On adaptive testing in orthogonal saturated designs. *Statistica Sinica*, in press.

Wang, W. and Voss, D. T. (2001). Control of error rates in adaptive analysis of orthogonal saturated designs. *Annals of Statistics*, **29**, 1058–1065.

Wang, W. and Voss, D. T. (2001). On the analysis of nonorthogonal saturated designs using effect sparsity. *Statistics and Applications* **3**, 177–192.

Wang, W. and Voss, D. T. (2003). On adaptive estimation in orthogonal saturated designs. *Statistica Sinica*, **13**, 727–737.

Wu, S. S. and Wang, W. (2004). Step-up simultaneous tests for identifying active effects in orthogonal saturated designs. Technical report 2004-018, *Department of Statistics, University of Florida.*

Ye, K. Q. and Hamada, M. (2000). Critical values of the Lenth method for unreplicated factorial designs. *Journal of Quality Technology*, **32**, 57–66.

Zahn, D. A. (1969). *An empirical study of the half-normal plot.* Ph.D. thesis, Harvard University, Boston.

Zahn, D. A. (1975a). Modifications of and revised critical values for the half-normal plots. *Technometrics*, **17**, 189–200.

Zahn, D. A. (1975b). An empirical study of the half-normal plot. *Technometrics*, **17**, 201–211.