Center for Quality and Productivity Improvement
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705

Report No. 16

STUDIES IN QUALITY IMPROVEMENT: ANALYSIS OF FRACTIONAL FACTORIALS

R. Daniel Meyer*

June 1986

* Current address: Lubrizol Corporation, 29400 Lakeland Boulevard, Wickliffe, OH 44092.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041, by the National Science Foundation Grant No. DMS-8420968, and by the Vilas Trust of the University of Wisconsin-Madison; and aided by access to the research computer of the University of Wisconsin Statistics Department. Completed August 1985.

CENTER FOR QUALITY AND PRODUCTIVITY IMPROVEMENT
UNIVERSITY OF WISCONSIN-MADISON

Report No. 16

STUDIES IN QUALITY IMPROVEMENT: ANALYSIS OF FRACTIONAL FACTORIALS

R. Daniel Meyer

June 1986

PRACTICAL SIGNIFICANCE

A problem frequently encountered in the industrial improvement of quality is to identify from among many variables, those which are responsible for large changes in the quality characteristics of a particular product or process. Statistically designed experiments, in particular fractional factorial designs, are a key tool in providing an economical solution to this problem.

This report contains a historical overview of the design and analysis of fractional factorials in the screening context described above, where the primary concern is to identify those experimental factors whose effects are too large to attribute to noise. Some new ideas are advanced for the problem of identifying these active factors, taking into account the notion of factor sparsity.

Key Words: Fractional factorial; screening experiment; unreplicated; active factors; effect sparsity.

Current address: Lubrizol Corporation, 29400 Lakeland Boulevard, Wickliffe, OH 44092.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041, by the National Science Foundation Grant No. DMS-8420968, and by the Vilas Trust of the University of Wisconsin-Madison; and aided by access to the research computer of the University of Wisconsin Statistics Department. Completed August 1985.

STUDIES IN QUALITY IMPROVEMENT: ANALYSIS OF FRACTIONAL FACTORIALS

R. Daniel Meyer

1. Introduction

The possible value of fractional factorial designs in industry seems to have been first recognized by Tippett (1934) (see also Fisher, 1966, p.88). To discover the cause of difficulty in a cotton-spinning machine, he succesfully screened five factors, each having five levels, in just 25 runs: a 125th fraction of a 5^5 factorial. A general framework for fractional factorials was described by Finney (1945). More general orthogonal array designs were introduced by Plackett and Burman (1946) and Rao (1947).

At the preliminary stages of an investigation, a two-level fractional factorial is very useful as a screening design. While Plackett and Burman (1946) gave a fairly complete enumeration of two-level designs involving a moderate number of runs, the 2^{k-p} fractional factorials are an especially useful subset and a thorough description of them was given by Box and Hunter (1961). Because the Hadamard product of any two columns of a 2^{k-p} design give another column of the design, the confounding structure is much simpler than for the general two-level orthogonal array. (The Hadamard product of two columns is defined to be a column with ith element equal to the scalar product of the ith elements of the original columns).

Current address: Lubrizol Corporation, 29400 Lakeland Boulevard, Wickliffe, OH 44092.

For ease of illustration, I will limit discussion here to two-level designs. It is assumed that the design matrix $\underset{\sim}{X}$ is a $n \times n$ orthogonal matrix of ± 1's such that $\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{X}\underset{\sim}{X}' = n\underset{\sim}{I}_n$, where $\underset{\sim}{I}_n$ is the $n \times n$ identity matrix. The first column $\underset{\sim}{x}_0$ of $\underset{\sim}{X}$ is a column of 1's, and one or all of the remaining columns $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_{n-1}$ are assigned to experimental variables; -1 denoting the low or nominal level and $+1$ denoting the high or alternate level. At the completion of the experiment the $n \times 1$ vector $\underset{\sim}{y} = (y_1, \ldots, y_n)'$ becomes available.

Typically, a linear model is employed for describing the observations from a two-level factorial experiment. At the screening stage of an investigation, it is often hoped that a first order model in main effects only will be adequate. This is written, with v the number of variables, as

$$\underset{\sim}{y} = \sum_{j=0}^{v} \underset{\sim}{x}_j \beta_j + \underset{\sim}{\varepsilon} \tag{1}$$

with the elements of the vector $\underset{\sim}{\varepsilon}$ assumed independently and normally distributed with zero mean and constant variance. (The main effect of variable j is usually defined to be twice the regression coefficient β_j.) If the above model were believed to be true, the parameters (including the error variance) could be efficiently estimated provided $(n - 1)-v$ was large enough to provide desired degrees of freedom for estimating the variance, or if repeat runs were included for this purpose. A model of this form would be adequate when the response was roughly planar over the experimental region examined. On the other hand, allowance should be made for the possible inadequacy of the model (1). Suppose the true response function was much closer to a second-order model of the form

$$\underset{\sim}{y} = \underset{\sim}{x}_0 \beta_0 + \sum_{j=1}^{v} \underset{\sim}{x}_j \beta_j + \sum_{i<j}^{v} (\underset{\sim}{x}_i \underset{\sim}{x}_j) \beta_{ij} + \underset{\sim}{\varepsilon} \tag{2}$$

where $\underset{\sim}{x}_i \underset{\sim}{x}_j$ is the Hadamard product of $\underset{\sim}{x}_i$ and $\underset{\sim}{x}_j$.

This would have the following implications. The estimate of the mean β_0 would be confounded with the pure quadratic coefficients β_{jj}. Estimates of the linear coefficients β_j may be confounded with interaction terms β_{ij}. The estimate of variance suplied by the $(n - 1)-v$ unassigned columns may also be biased by real interaction effects.

To guard against the problems outlined above, one could take several approaches.

A second-order design could be employed which allowed estimation of all parameters of the model (2) (see, e.g., Box and Hunter, 1957). However, this greatly reduces the number of factors which could be studied in a given number of experimental runs.

The inclusion of replicate runs in the two-level design would allow unbiased estimation of the variance. Lack of fit of the model (1) could be detected by the presence of large contrasts associated with the $(n - 1)-v$ unassigned columns, and the design could be augmented to estimate the full second-order model, if necessary (Box and Wilson, 1951). However, the requirement of replicate runs again reduces the number of factors which could be studied in a given number of runs.

A third approach relies on a phenomenon of "effect sparsity" (Box and Meyer, 1985). The object of a screening experiment is to isolate important factors among a group of many candidates. If this is possible, then even if the true response was more closely approximated by the second-order model (2), many of the parameters would be negligible compared to the parameters associated with the important variables and the effect of noise. In this case an unreplicated two-level design will yield $n - 1$ estimated effects, most of which will be inert and attributable to noise, the remainder of which will be active and too large to attribute to noise. As above, inadequacy of the

first-order model could be detected by the presence of a large contrast associated with one of the $(n-1)-v$ unassigned columns.

This last approach, while combining the virtues of relatively low cost and relatively great information, does not always supply unambiguous results. Confounding of effects may lead to more than one plausible explanation of the data. However, a follow-up experiment to resolve ambiguities would usually involve fewer variables and many fewer runs than the original experiment, and the combined cost would be less than the cost of a completely comprehensive experiment in all variables (Box, Hunter and Hunter, 1978).

2. Analysis of Fractional Factorials

Analysis of fractional factorial experiments has traditionally involved, primarily, identifying and estimating the active effects. In addition, estimating the error variance may also be of interest. The process of identifying the active effects has historically been divided into two stages (see, e.g., Box, Hunter and Hunter, 1978, Chapter 12). The first stage involves identifying the orthogonal contrasts $T_i = x_i'y/n$, $i = 1, \ldots, n-1$, which are too large to be attributed to noise. These are called active contrasts. Under the second-order model (2) the expected value of T_i will be a linear combination of one or more of the coefficients β, sometimes called an alias string when involving more than one parameter. Under the hypothesis of effect sparsity, however, most of the contrasts will have expectation zero. A small proportion will have active terms in their alias string, and these will have non-zero expectation. The second stage of the analysis then involves determining which of the experimental factors are associated with the active contrasts.

2.1. First stage analysis

Some of the techniques which have been employed to identify active contrasts are as follows:

Analysis of variance has been used to judge the reality of the contrasts (see Davies ed., 1954, p. 464). This method relies on comparison of the contrasts with an independent measure of error variance. When an estimate of experimental error variance is available from relevant genuinely replicated runs from current or past experimentation, construction of the analysis of variance table is straightforward.

For unreplicated experiments, it has been customary to identify a priori certain contrasts, usually those which have only higher order interactions in their alias strings, whose magnitude could be attributed solely to random error. (In the case of quantitative factors, relative smoothness of the response surface would dictate that higher order interactions, which correspond to higher order derivatives, become successively smaller. This is reasonable as long as the ranges for the variables are chosen moderately. Likewise, for qualitative variables, the existence of higher order interactions implies a wide difference between levels of the variables, which should be avoided. Alternately, if the levels of qualitative variables must be chosen to be very dissimilar, separate experiments should be run for each level. In this way the frequency of large, high order interactions can be minimized, and contrasts which measure these interactions can be assumed to measure noise.) These inert contrasts are then used to estimate error variance. This approach necessarily restricts the degree of fractionation to be used in the design, as several columns must be reserved to estimate effects supposedly known to be inert. Alternately when little is known about which effects are inert, the required contrasts may be difficult or impossible to

identify. An even less satisfactory procedure for estimating the experimental error variance employs successive pooling of supposedly nonsignificant components in the analysis of variance table.

Daniel (1959) introduced the half-normal plot for judging the significance of orthogonal contrasts from a factorial experiment. In this method the $n - 1$ ordered absolute contrasts $|T_{(i)}|$ are plotted against $\Phi^{-1}(1/2 + (i - 1/2)/2(n - 1))$, where Φ is the standard normal distribution function. Under the completely null hypothesis of no active contrasts, these points should fall roughly along a straight line through the origin. Contrasts too large to be explained by noise would appear as extreme points falling off the line. Later, Daniel (1976) pointed out that any information contained in the signs of the contrasts is obscured in the half-normal plot. A slight modification of Daniel's idea, the full normal plot, i.e., plotting the signed ordered contrasts $T_{(i)}$ against $\Phi^{-1}((i - 1/2)/(n - 1))$, can be interpreted in the same way as the half-normal plot without losing the diagnostic information in the signs of the contrasts.

The advantages of normal probability plotting are that it requires neither replicated runs nor prior identification of inert contrasts and also allows for selection automatically. As with other graphical procedures, the normal plot may suggest further examination of the data. In particular, it can be used to detect model inadequacies.

Daniel also suggested how formal inference about which contrasts were significantly non-zero could be implemented through the normal plot. "Guardrails" of various Type I error rates are constructed by considering the null distributions of the ordered absolute contrasts. In a companion paper, Birnbaum (1959) discussed several methods for judging which contrasts measured non-zero effects, and showed that Daniel's procedure could be regarded as an

approximation to the optimal statistic when there was at most one significant contrast. In addition, Birnbaum stated that the optimal procedure for the case of more than one significant contrast was far too complicated for practical application, and concluded that Daniel's analysis was preferable for typical research applications. Zahn (1975) proposed some revisions to Daniel's procedure, including corrections to the critical values of the test.

Two other methods for analyzing unreplicated factorials were given by Wilk, Gnanadesikan and Freeny (1963) and Holms and Berrettoni (1969). Wilk, et al. suggested using maximum likelihood estimation of the variance, assuming that some number K of the original contrasts only measure error. The estimation is then based on the M ($<K$) smallest contrasts in order to avoid including contrasts measuring real effects in the estimate of σ, with suggested choice of M being $0.7K$. However, their estimate of σ was shown to be quite sensitive to the choice of K. Holms and Berrettoni proposed a method for the case when it is expected that a large proportion of the contrasts measure real effects. They considered the ordered absolute contrasts from smallest to largest, with each one in turn compared to those smaller than it. Critical values of the procedure, called "chain-pooling," were derived from work done by Cochran (1941).

2.2. Second stage analysis

Box and Hunter (1961) offered two guidelines for the process of associating factorial effects with active contrasts in the presence of confounding:

1. Main effects are more likely to occur than two-factor interactions, which are more likely than three-factor interactions, etc. That is, if a large contrast is associated with more than one effect, the effect of lowest order is usually considered the most likely cause. This is especially

true for continuous variables, when smoothness of the response surface dictates that higher-order effects, which correspond to higher-order derivatives, become successively smaller. In screening situations and other applications, it is common to ignore three-factor or higher order interactions.

2. Variables which have large main effects are more likely to have significant interactions among themselves or with other variables. For example, when a large contrast is associated with several two-factor interactions, the interactions involving variables with large main effects are considered more likely to be the cause.

The authors emphasize that these guidelines are to be employed to make tentative conclusions, subject to verification by subsequent experimentation or monitoring of the process after implementing changes. Exceptions to the rules appear, for example, when the design is located on a diagonal ridge of the response surface. This can occur when the process has been fine-tuned in the past one variable at a time, in the presence of compensating factors such as time and temperature of a chemical reaction. The experiment will then produce small main effects among the compensating factors, but a large two-factor interaction.

3. A Bayesian Approach

The assumptions that are made when analyzing factorials and fractional factorials can be modeled formally. Once the assumptions are made explicit, Bayes' theorem provides a straightforward method of inference.

3.1. Identification of active contrasts

To model the assumption that a majority of the column contrasts are expected to be inert, it is assumed there is some prior probability α that

each column is active, with α generally assumed to be less than 1/2. Let $a_{(c)}$ denote the event that a particular combination of c of the $n - 1$ contrasts are active, the remainder inert. The prior probability of the event $a_{(c)}$ is

$$p(a_{(c)}) = \alpha^c(1 - \alpha)^{n-1-c} .$$

After observing the data y from the experiment, the posterior probability of the event $a_{(c)}$ is

$$p(a_{(c)}|y) = \frac{p(y|a_{(c)})p(a_{(c)})}{\sum_{(i)} p(y|a_{(i)})p(a_{(i)})} , \tag{3}$$

where the denominator is the summation over all possible combinations of active and inert columns, and $p(y|a_{(c)})$ is the predictive density of the observations y given $a_{(c)}$. Of particular interest is the marginal posterior probability that column i is active, and this is given by

$$p_i = P[\text{column } i \text{ active}|y] = \sum_{(c):i \text{ active}} p(a_{(c)}|y) . \tag{4}$$

Inference about which columns are active can be made from the probabilities $\{p_i\}$.

3.2. Identification of active factors

Once active columns have been identified, it remains to identify which factors are responsible for the large contrasts. Alternately, in some situations there may only be interest in which factors are active, regardless of how their activity can be explained by main effects, interactions, etc. A modification of the Bayesian model introduced above is useful for this type of analysis.

Rather than contrasts being active with some prior probability α, it is assumed that factors will be active with probability α, with a suitable

adjustment in the value of α. The notation $a_{(f)}$ would now refer to the event that a particular combination of f factors (including possible interactions) was active. The posterior probability of $a_{(f)}$ is then derived analagous to the expression for the posterior probability of $a_{(c)}$ given previously. The posterior probability of each factor being active is then

$$p_i^* = \sum_{(f):i \text{ active}} p(a_{(f)}|y) \ . \tag{5}$$

4. Bad Values in Fractional Factorials

As noted by Daniel (1976) and others, the results of unreplicated fractional factorial experiments are sensitive to bad values among the observations. Daniel (1959) estimated that in his experience, the relative frequency of bad values in factorial experiments was anywhere from .01 to .1, depending on the complexity of the experimental situation and on the experience of the experimenter. If, for example, each observation in a 16-run experiment had an independent probability of .05 of being incorrectly determined, then over half of such experiments would contain one or more bad values. Daniel also felt that quite often the presence of large higher-order interactions in factorial experiments was not due to highly curved response surfaces, but to erroneous observations which were not identified as such. Because of the saturated nature of unreplicated factorial experiments, bad observations can often be concealed by mistaken identification of some combination of active effects.

Full normal plotting of the observed contrasts (Daniel, 1959, 1976) has been a useful diagnostic tool for detecting bad values in unreplicated experiments, in addition to its use in identifying active contrasts. If a particular observation is biased by, say, a positive amount, those contrasts in which the observation enters positively are shifted to the right, and those

contrasts in which the observation enters negatively are shifted to the left. This produces a "gap" among the inert contrasts of the normal plot which is the telltale sign of a bad observation. Similarly, the presence of multiple bad values can produce multiple gaps in the normal plot.

There is a wide literature on the general statistical issue of outliers. In their review article, Beckman and Cook (1983) list 229 references concerning the detection and accommodation or rejection of bad values. While some general regression diagnostics could conceivably be employed in the analysis of factorial experiments, Little (1983) for example showed that several of the common diagnostics consisted of a factor which measured the leverage of the doubtful points in the X space and a factor which measured the change in the residual sum of squares when suspected bad values were deleted. Because all design points have equal leverage in two-level factorials, for the case of one outlier the leverage factor would be the same for each observation, and in general one would not expect to have problems with extreme points in the X space from factorial experiments. Thus these diagnostics would reduce to functions of the change in the residual sum of squares when bad values are deleted, and some methods, described below, have been developed for factorial designs which are basically functions of the change in the residual sum of squares.

Daniel (1960) proposed a test for bad values based on the maximum residual after active contrasts have been identified. The observation corresponding to the maximum residual is identified as bad if the modulus of the residual is greater than a specified upper percentile of its null distribution. Stefansky (1972) derived revised critical values for Daniel's test.

John (1978) described a general method for detecting one or two bad values in a factorial experiment, based on work by Gentleman and Wilk (1975a,b), John and Prescott (1975a,b), and John and Draper (1978), which incorporates the reduction in the sum of squares when supposed bad values are deleted. The method is similar to one proposed by Goldsmith and Boddy (1973), and encompasses Daniel's test based on the maximum residual.

The existing methodology for dealing with bad values in unreplicated factorials supposes that a fixed model has been identified. However, the possibility of bad values may also be accommodated by "robustifying" the sampling distribution of y. Specifically, it is assumed the errors in the model (1) come from the scale-contaminated normal distribution denoted by

$$(1 - \alpha_2)N(o,\sigma_\varepsilon^2) + \alpha_2 N(0,k_2^2\sigma_\varepsilon^2)$$

(Jeffreys, 1932; Dixon, 1951; Tukey, 1960; Box and Tiao, 1968). This provides for assessment of the contrasts while simultaneously allowing for the possibility of bad values, to be contrasted with the practice of checking for bad values after the active contrasts have been identified. Using the Bayesian approach above, the posterior probability that a contrast is active can be calculated while taking into account the possibility of bad values, and the posterior probability that an observation is "bad" can be calculated in light of the consideration that the identity of the active contrasts is not known. In this way questionable observations can be identified and investigated, and the sensitivity of the conclusions to the presence of possible bad values can be measured.

The robustification of models in this way, while philosophically attractive, has been historically difficult to implement because it usually requires very extensive computing. However, as the speed and sophistication of computers has advanced, computationally intensive statistical analysis has

become more feasible. At present the amount of computing time needed to analyze the above model allowing for bad values in full generality is not practical. Various computational shortcuts may be used, however, based on reasonable assumptions about the maximum possible number of bad values and active contrasts. It seems likely that future advances in computing technology will reduce such limitations.

5. Posterior Probabilities for Model Selection

For the general problem of model discrimination, Atkinson (1978) presents three objections to the use of posterior probabilities:

1. When all of the candidate models fit badly, the best fitting of these will be chosen with high probability for n large enough.
2. When competing models have different numbers of parameters, the model with fewest parameters is favored in the absence of evidence in the data.
3. When models are nested and the simplest model is true, then all models are true and should receive equal weight. However, the simplest model will receive the highest posterior probability (by argument 2 above), and the remaining models will receive decreasing weights depending on the number of parameters.

The Bayesian analysis proposed above could be classified as a model discrimination procedure. The objections listed above as they relate to the proposed analysis, are answered as follows:

1. Any statistical estimation procedure chooses the best-fitting model from a family of models according to some criterion. Usually this involves estimating parameters which index the family of models. However, there ought to be no implication in such a procedure that the best-fitting model among those considered will be adequate. Likewise, a model

identified by a high posterior probability need not fit well. Diagnostic model-checking is an essential part of any statistical analysis (see Box, 1980).

2. The validity of this objection for the general question of model discrimination will not be discussed here. Under the assumed condition of effect sparsity, the favoring of a model with fewer parameters, in the absence of evidence from the data, may be viewed as an advantage rather than a disadvantage.

3. When a model is "true", the statement that "all models in which the true model is nested are also true" is a matter of philosophy. One might argue that there is an inherent difference between the models

$$M_1 : Y = \theta_1 X_1 + 0(X_2)$$

and

$$M_2 : Y = \theta_1 X_1 + \theta_2 X_2 .$$

In any event, the result that the simpler model receives higher posterior probability does not seem misleading.

6. Summary

Unreplicated factorial designs have been and still are a valuable tool in industrial experimentation, despite the fact they do not allow for the estimation of error variance usually obtained from repeat runs. Methods of analysis used in the past have depended more or less on an implicit assumption about the sparsity of real effects. If such assumptions are explicitly incorporated into the usual linear model employed for such experiments, inference about active and inert contrasts is more straightforward, and dependence of the inference on the prior assumptions is more easily assessed.

It is assumed that there is a prior probability α that each of the orthogonal contrasts is active, i.e., measures a real effect, and contrasts

are active independently of one another. Assuming a normal prior distribution for the expected value of an active contrast, the posterior probability that a contrast is active can be computed. While computations of this sort generally require extensive computing time, an alternative Bayes factorization allows the posterior probabilities to be obtained by numerical integration at a considerable reduction in computing requirements.

Further assumptions about the size and relative frequency of main effects and interactions can be incorporated into the model. It is shown by Meyer (1985) that the posterior probabilities that experimental factors are active combines prior assumptions, properties of the design and information in the data. Factors which can not be safely eliminated as inert due to the confounding pattern of the design will receive significant posterior probability in addition to those factors which are more obviously active.

The model can be extended to allow for the possibility of discordant observations. Such observations are assumed to have inflated variance with prior probability α_2. Given this model, the posterior probability that a contrast (or factor) is active could be computed taking into account the possibility of bad observations, as well as the probability that a particular observation is bad, i.e., has inflated variance. It is shown by Meyer (1985) that the approach of testing residuals for outliers after active contrasts are identified is sometimes inferior to the Bayesian model-based approach.

The extension to the possibility of bad observations greatly increases the computing requirements of the analysis, so that they are often unfeasibly high. An iterative analysis is proposed by Meyer (1985) as an exploratory method rather than a numerical approximation. A method of approximating the posterior probabilities which possessed good numerical properties would be one area of future research.

Acknowledgement

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041, by the National Science Foundation Grant No. DMS-8420968, and by the Vilas Trust of the University of Wisconsin-Madison; and aided by access to the research computer of the University of Wisconsin Statistics Department. Completed August 1985.

References

Atkinson, A. C. (1978) Posterior Probabilities for Choosing a Regression Model. Biometrika, 65, 39-48.

Beckman, R. J. and Cook, R. D. (1983) Outlier...s. Technometrics, 25, 119-164.

Birnbaum, Allan (1959) On the Analysis of Factorial Experiments Without Replication. Technometrics, 1, 343-357.

Box, George E. P. and Wilson, K. B. (1951) On the Experimental Attainment of Optimum Conditions. J. R. Statist. Soc. B, 13, 1-45.

Box, G. E. P. and Hunter, J. S. (1957) Multi-factor Experimental Designs for Exploring Response Surfaces. Ann. Math. Statist., 28, 195-241.

Box, George E. P. and Hunter, J. Stuart (1961) The 2^{k-p} Fractional Factorial Designs. Technometrics, 3, 311-351, 449-458.

Box, George E. P. and Tiao, George C. (1968) A Bayesian Approach to Some Outlier Problems. Biometrika, 55, 119-129.

Box, George E. P., Hunter, William G. and Hunter, J. Stuart (1978) Statistics for Experimenters. New York:John Wiley.

Box, George E. P. (1980) Sampling and Bayes' Inference in Scientific Modelling and Robustness. J. R. Statist. Soc. A, 143, 383-430.

Box, G. E. P. and Meyer, R. D. (1985) Analyzing Two-Level Fractional Factorial Experiments for Possible Dispersion Effects. TSR #2746, Mathematics Research Center, University of Wisconsin-Madison.

Cochran, W. G. (1941) The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total. Annals of Eugenics, **11**, 47-52.

Daniel, Cuthbert (1959) Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments. Technometrics, **1**, 311-341.

Daniel, C. (1960) Locating Outliers in Factorial Experiments. Technometrics, **2**, 149-156.

Daniel, Cuthbert (1976) Applications of Statistics to Industrial Experimentation. New York:John Wiley.

Davies, O. L., ed. (1954) Design and Analysis of Industrial Experiments. London:Oliver and Boyd.

Dixon, W. J. (1951) Ratios Involving Extreme Values. Ann. Math. Statist., **22**, 68-78.

Finney, D. J. (1945) The Fractional Replication of Factorial Arrangements. Annals of Eugenics, **12**, 291-301.

Fisher, R. A. (1966), The Design of Experiments, 8th edition. New York:Hafner.

Gentleman, J. F. and Wilk, M. B. (1975) Detecting Outliers in a Two-Way Table: I. Statistical Behaviour of Residuals. Technometrics, **17**, 1-14.

Gentlemen, J. F. and Wilk, M. B. (1975) Detecting Outliers: II. Supplementing the Direct Analysis of Residuals. Biometrics, **31**, 387-410.

Goldsmith, P. L. and Boddy, R. (1973) Critical Analysis of Factorial Experiments and Orthogonal Fractions. Appl. Statist., **22**, 141-160.

Holms, A. G. and Berrettoni, J. N. (1969) Chain-Pooling ANOVA for Two-Level Factorial Replication-Free Experiments. Technometrics, **11**, 725-746.

Jeffreys, H. (1932) An Alternative to the Rejection of Observations. Proceedings of the Royal Society, London, A, **137**, 78-87.

John, J. A. and Prescott, P. (1975) Critical Values of a Test to Detect Outliers in Factorial Experiments. Appl. Statist., **24**, 56-59.

John, J. A. and Prescott, P. (1975) Estimating Missing Values in Experiments. Appl. Statist., **24**, 190-192.

John, J. A. (1978) Outliers in Factorial Experiments. Appl. Statist., **27**, 111-119.

John, J. A. and Draper, N. R. (1978) On Testing for Two or One Outliers in Two-Way Tables. Technometrics, **20**, 69-78.

Little, J. K. (1983) Analysis of the Scale-Contaminated Normal Model: Diagnostics and Robustness. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

Meyer, R. D. (1985) A New Approach to the Analysis of Fractional Factorials. Report #2865, Mathematics Research Center, University of Wisconsin-Madison.

Plackett, R. L. and Burman, J. P. (1946) Design of Optimal Multifactorial Experiments. Biometrika, **23**, 305-325.

Rao, C. R. (1947) Factorial Experiments Derivable from Combinatorial Arragements of Arrays. JRSS, **B9**, 128-140.

Stefansky, W. (1972) Rejecting Outliers in Factorial Designs. Technometrics, **14**, 469-479.

Tippett, L. H. C. (1934) Applications of Statistical Methods to the Control of Quality in Industrial Production. Manchester Statistical Society.

Tukey, J. W. (1960) A Survey of Sampling from Contaminated Distributions. Essays in Honor of Harold Hotelling, J. Olkin et al. (eds.), Stanford University Press.

Wilk, M. B., Gnanadesikan, R. and Freeny, A. E. (1963) Estimation of Error Variance from Smallest Ordered Contrasts. J. Amer. Statist. Ass., **58**, 152-160.

Zahn, Douglas A. (1975) Modifications of and Revised Critical Values for the Half-Normal Plot. Technometrics, **17**, 189-200.