Statistica Sinica **8**(1998), 1-41

ANALYZING UNREPLICATED FACTORIAL EXPERIMENTS: A REVIEW WITH SOME NEW PROPOSALS

M. Hamada and N. Balakrishnan

University of Michigan and McMaster University

Abstract: Recently, there have been many proposals for objectively analyzing unreplicated factorial experiments. We review these methods along with some earlier and perhaps lesser known ones. New methods are also proposed. The primary aim of this paper is to compare these methods and their variants via an extensive simulation study. Robustness of the various methods to non-normality is also considered. Many methods are comparable, but clearly some cannot be recommended. The results from the study also suggest some basic principles for evaluating new methods. Finally, we outline some issues that this study has raised and which might benefit from work in other areas such as multiple comparisons, outlier detection, ranking and selection, and robust statistics.

Key words and phrases: Bayesian, censoring, correlation coefficient, half-normal probability plot, interquartile range, mean squares, MLE, order statistics, outlier, pooling robust statistics, sequential procedure, Shapiro-Wilk test, trimming, variance homogeneity.

1. Introduction

Since the 1980's, the objective analysis of unreplicated two-level factorial and fractional factorial designs has attracted much attention. The analysis of unreplicated experiments with say n runs presents a challenge because while $n-1$ effects (excluding the overall mean) can be estimated by contrasts, there are no degrees of freedom left to estimate the error variance. Consequently, standard t tests cannot be used to identify the "active" effects.

In practice, the standard method for identifying active effects continues to be a probability plot of the contrasts, the first method for this problem proposed by Daniel (1959). See Daniel (1983) for an interesting personal recollection. Plotting the unsigned contrasts on half-normal probability paper, the contrasts for the "inert" effects fall along a straight line while those for the active ones tend to fall off the line. There is a subjective element in deciding what constitutes "falling off the line", which has motivated the recent work to provide an objective method.

This paper reviews various methods for analyzing unreplicated experiments given in Box and Meyer (1986), Voss (1988), Lenth (1989), Benski (1989), Bissell

(1989, 1992), Berk and Picard (1991), Juan and Pena (1992), Loh (1992), Le and Zamar (1992), Dong (1993), Schneider, Kasperski and Weissfeld (1993) and Venter and Steel (1996). This flurry of activity seems to have been motivated in part by Taguchi's practice (Taguchi (1987)) of pooling the smallest contrasts to estimate the error variance in an ANOVA and applying the usual F distribution critical values (Box (1988), Bissell (1989) and Berk and Picard (1991)). The methods proposed in two lesser known papers, Seheult and Tukey (1982) and Johnson and Tukey (1987), are also studied as well as earlier work by Holms and Berrettoni (1969) and Zahn (1975a, b). Note that Daniel's (1959) proposal did provide an objective method, guardrails on a standardized half-normal plot, but for the most part has been ignored. In order to analyze unreplicated experiments, the assumption that at least some of the effects are "inert" needs to be made. In fact, most of the existing methods assume effect sparsity, that only a few effects are active, which seems to hold up in practice, say 20% (Box and Meyer (1986)). Daniel (1976), p. 75 suggested 25% and lowered it to 20% in Daniel (1983). These methods have varied motivations which will be considered in more detail in Section 2.

In this paper, we focus on methods based on the unsigned contrasts or their corresponding mean squares. This is because of the arbitrariness of "low" and "high" factor level labels which has been pointed out by Shapiro and Wilk (1965), Seheult and Tukey (1982) and Loh (1992). Since the method's results should not depend on the labeling, we consider the half-normal version, e.g., the half-normal probability plot of the unsigned contrasts rather than the normal probability plot of the contrasts. Note that Daniel (1976, 1983) prefers the normal probability plot for detecting problems with the data such as outliers, however.

We need a common notation to resolve some conflicts in the literature. Lists of the notation used and methods studied in this paper are provided below for easy reference. There are k $(= n - 1)$ effects denoted by κ_i, $i = 1, \ldots, k$, e.g., 7, 15, 31 for the commonly run 8, 16 and 32 run designs. By contrast or estimated effect, we mean the difference of the averages of the observations at the high and low levels; the contrasts are denoted by c_i and the unsigned contrasts by $|c_i|$. Also, the ith ordered unsigned contrast out of j contrasts is denoted by $|c|_{(i)j}$. The process or error variance is σ^2, so that the error variance of c_i denoted by τ^2 is $[4/(k+1)]\sigma^2$. Thus the problem is to decide which κ_i are active, i.e., non-zero, using the contrasts c_i. Normally distributed errors are assumed so that the contrasts c_i are normally distributed. Finally, the size of the active effect κ_i will be given in multiples of σ, the process or error standard deviation.

In Section 4, the paper compares the methods listed above as well as some new proposals presented in Section 3. To date, only limited studies comparing some of these methods have been done: Zahn (1975b), Voss (1988), Berk and

Picard (1991), Loh (1992), Dong (1993), Haaland and O'Connell (1995), Benski (1995) and Benski and Cabau (1995). The problem in comparing these methods is to do so on an equitable basis. For example, the methods perform differently when all the effects are inert. Thus, the methods need to be calibrated as much as possible without destroying the essence of the methods. The goal of the comparison is to identify the good performers. Performance of the procedures is evaluated through an extensive simulation study, which includes a limited investigation into their performance under nonnormality.

List of Notation

n	number of runs
k	number of contrasts
k'	current number of contrasts being considered in sequential test
σ	error standard deviation
τ	contrast standard error
κ_i	ith effect
c_i	ith contrast - estimate of ith effect
$\lvert c\rvert_{(i)}$	ith order statistic of the unsigned contrasts
$\lvert c\rvert_{(i)k}$	ith order statistic out of k unsigned contrasts
M_i	ith mean square
$M_{(i)}$	ith smallest mean square
$\lvert z\rvert_{(i)j}$	expected value of ith standard half-normal order statistic out of j
$\lvert\tilde{z}\rvert_{(i)j}$	median of ith standard half-normal order statistic out of j
$z_{(i)j}$	expected value of ith standard normal order statistic out of j
$\tilde{z}_{(i)j}$	median of ith standard normal order statistic out of j
α_{pool}	pooling level in Holms and Berrettoni (1969)
α_{final}	final level in Holms and Berrettoni (1969)
floor$[x]$	largest integer less than x
ceiling$[x]$	smallest integer greater than x
d_F	interquartile range
α_{active}	probability of an effect being active (Box and Meyer 1986)
K	model parameter for active effects in Box and Meyer (1986)
PSE	pseudo standard error; see Lenth (1989) (11)
IMAD_0	iterated median absolute deviation; see Juan and Pena (1992)
p_i	probability of declaring i effects active under all inert effects
EER	experimentwise error rate
IER	individual error rate
#AE	number of active effects

In Section 5, the paper presents a basis for evaluating new methods and lists some issues raised by the study that merit further attention. Section 6 concludes with some specific recommendations for the practitioner.

List of Methods

BEN89	Benski (1989)
BI89	Bissell (1989) uses (12)
BI92	Bissell (1992) uses (14)
BM86	Box and Meyer (1986)
BP91	Berk and Picard (1991) uses (13)
CORR	correlation coefficient probability plot uses (26)
DAN59	Daniel (1959) uses (1)
DISP	dispersion test uses (29)
DONG93	Dong (1993) uses (22)
HB69	Holms and Berrettoni (1969) uses (2)
HLOH92	half-normal version of Loh (1992)
HSW	half-normal Shapiro-Wilk uses (28)
JP92	Juan and Pena (1992) uses (21)
JTUK87	Johnson and Tukey (1987) uses (6)
LEN89	Lenth (1989) uses (11)
MDONG	uses iterative DONG93 estimator
MLEN	LEN89 accounting for k'
MLZ92	modified Le and Zamar uses (18)
MSKW	SKW93 accounting for k'
SKW93	Schneider et al. (1993) uses (23)
STUK82	Seheult and Tukey (1982)
VS96	Venter and Steel (1996) uses (25)
ZAHN75	Zahn (1975a, version S) uses (4)
ZAHN75(m)	uses m smallest contrasts to estimate τ

2. Existing Methods

Daniel (1959)

Daniel (1959) used the idea of detecting outliers in a data set by probability plotting as discussed above: i.e., the outliers, those falling off the line, correspond to active effects. Note the implicit assumption of few active effects in order to draw a line through bulk of small contrasts and how this method ingeniously avoids the need for estimating σ. Its subjectivity was mentioned above, however.

Daniel (1959) also presented an objective graphical method, a standardized probability plot with guardrails, which plots the unsigned contrasts divided by the ordered unsigned contrast corresponding to order statistic closest to the 0.683

percentile. Note that the 0.683 percentile of the half-normal distribution is equal to τ, and suggests an estimate for the contrast standard error τ when all the effects are inert. Thus, for example, for $k = 15$ effects, the unsigned contrasts are standardized by $|c|_{(11)}$ and have the form:

$$|c|_{(i)}/|c|_{(11)}. \tag{1}$$

These statistics are referred to as *modulus ratios* since they are ratios of modulii or absolute values. Note how the unknown scale is removed by the standardization and that only about 25% of the largest unsigned contrasts can be tested sequentially starting with the largest. The guardrails are the corresponding critical values drawn on the plot, where the critical value for $|c|_{(i)}$ is based on the distribution of $|c|_{(i)k}$. Active effects are then identified by the standardized contrasts which exceed their corresponding guardrails. Birnbaum (1959) gave approximations for the distribution of the largest modulus ratio and showed that it is the most powerful test when there is only one active effect.

Holms and Berrettoni (1969)

Holms and Berrettoni (1969) proposed a method called *chain-pooling*. The method works with the mean squares M_i which are proportional to the squared contrasts c_i^2 and compares the largest standardized mean squares. The standardization is based on the smallest mean squares whose corresponding effects are likely not active; the determination whether a particular mean square is pooled or not is based on all smaller mean squares.

More formally, starting with the m (possibly equal to one) smallest mean squares, use $U_{(m+1)} = (m+1)M_{(m+1)}/\sum_{i=1}^{m+1} M_{(i)}$ to determine whether the next largest mean square $M_{(m+1)}$ should be pooled or not at level α_{pool}, say 0.25. Pooling is stopped once the p-value falls below α_{pool}. Then declare active those effects corresponding to the larger mean squares whose p-values are less than α_{final} using:

$$jM_{(l)}/\Big(\sum_{i=1}^{j-1} M_{(i)} + M_{(l)}\Big), \tag{2}$$

where the $j-1$ smallest M_i's are pooled. Critical values based on all inert effects for $k = 15$ are given in their Table 1. That is, α_{pool} controls how many of the smallest mean squares are pooled while α_{final} controls how many of the largest mean squares are declared significant. Thus, a strategy is defined by m, α_{pool} and α_{final}. The motivation for this procedure was the case when there are a large number of active effects; thus, an estimate for error variance needs to be based on a small number of contrasts in which case m should be set small and possibly to one.

Zahn (1975a)

Motivated by Daniel (1959), Zahn (1975a) proposed using an alternative estimate of the contrast standard error for standardizing the unsigned contrasts based on 68.3% of smallest unsigned contrasts. That is, τ can be estimated by the slope of the regression line through the origin on Daniel's half-normal plot:

$$S_{\text{ZAHN}} = \sum_{i=1}^{m} |c|_{(i)} |z|_{(i)k} / \sum_{i=1}^{m} |z|^2_{(i)k}, \tag{3}$$

where $m = \text{floor}[0.683k + 0.5]$ and $|z|_{(i)k}$ is the expectation of $|c|_{(i)}$. Zahn (1975b) showed that S_{ZAHN} has a smaller mean squared error (MSE) than $|c|_{(11)}$ for $k = 15$ which explains the suggestion, his Version S, of using:

$$|c|_{(i)} / S_{\text{ZAHN}}. \tag{4}$$

Like (1), (4) is designed for testing only a few of the largest unsigned contrasts, i.e., four for $k = 15$ since $m = 11$. In contrast with Daniel (1959), the critical value for $|c|_{(i)}$ is based on the distribution of $|c|_{(i)i}$, i.e., the ith or largest order statistic in a sample of size i.

Zahn (1975b) also studied Versions XR and SR based on (1) and (4) respectively, where τ is re-estimated in subsequent tests based on a variable m, $m' = \text{floor}[0.683k' + 0.5]$, where k' is the current number of contrasts being considered. Note that $|z|_{(i)k'}$ is used in estimating τ in both the S and SR versions.

Seheult and Tukey (1982)

Seheult and Tukey (1982) used an outlier procedure based on the quartiles of a synthetic batch of contrasts, namely zero plus all the contrasts with both signs giving a total of $2^{k+1} - 1$ items. The threshold is twice the interquartile range or, because of the symmetry of the synthetic batch, is four times the median of the unsigned contrasts plus zero. In the terminology coined by Tukey (1977), the outliers are those exceeding one-and-a-half hinge spreads outside the nearest hinge. Assuming normality, the probability of exceeding the threshold is very small, 0.007. Seheult and Tukey (1982) then proposed using this threshold iteratively by removing the largest contrast and its associate if they exceed the threshold and applying the procedure to the remaining $2^k - 1$ synthetic contrasts and so on.

Box and Meyer (1986)

Box and Meyer (1986) presented a Bayesian approach based on effect sparsity, i.e., there is a small proportion of active effects α_{active}. They used a scale contaminated model which assumes that the active effects κ_i have a $N(0, \sigma^2_{active})$ distribution. Thus, contrasts c_i corresponding to active effects have distribution

$N(0, K^2\sigma^2_{inactive})$, where $K^2 = (\sigma^2_{inactive} + \sigma^2_{active})/\sigma^2_{inactive}$; contrasts c_i corresponding to inert effects have distribution $N(0, \sigma^2_{inactive})$. For each effect, the marginal posterior probability of being active is computed and declared active if the probability exceeds 0.5. Specifically, the posterior probability of each of the possible 2^k models (i.e., an effect is active or not) is first computed. Then, the marginal posterior probability is the sum of the posterior probabilities over all those models containing the particular effect. Box and Meyer (1986) noted that estimates for α_{active} and K based on ten published analyses of data sets was (0.13-0.27) and (2.7-18) with averages of 0.2 and 9.6, respectively. This provides empirical support for the principle of effect sparsity and motivated their recommendation of 0.2 and 10 for α_{active} and K, respectively.

Johnson and Tukey (1987)

Johnson and Tukey (1987) proposed a procedure based on display ratios which are the unsigned contrasts divided by their respective typical order statistics; i.e.,

$$|c|_{(i)}/|\tilde{z}|_{(i)k}, \tag{5}$$

where $|\tilde{z}|_{(i)k}$ is the median of the half-normal ith order statistic in a sample of size k. Their motivation for the display ratios was to make comparison easier since the natural reference line is now horizontal with its height being an estimate of τ. Contrast this with the half-normal plot, whose natural reference line is a line through the origin whose slope is an estimate of τ.

The objective method that Johnson and Tukey (1987) proposed is based on ratio-to-scale statistics which are computed as:

$$\text{ratio-to-scale} = \text{display ratio/median display ratio}. \tag{6}$$

Critical values for the ith largest ratio-to scale statistic given in their Table 12 are for the ith largest or maximum ratio-to-scale statistic in a sample of size i. Johnson and Tukey (1987), p. 203 then proposed using the ratio-to-scale statistics sequentially, dropping the contrast corresponding to the maximum ratio-to-scale and applying the procedure to the remaining contrasts. Note the similarity with Daniel (1959) except that display ratios are used and the denominator is the median rather than the 0.683 percentile.

Voss (1988)

Voss (1988) presented what he termed generalized modulus ratio (GMR) tests. He considered non-decreasing functions f of the $|c|_{(i)}$ standardized by a linear combination of them:

$$f(|c|_{(i)})/\sum a_i f(|c|_{(i)}), \tag{7}$$

for some constants a_i. Note that (1), (2), (4) and (13) to be discussed later fall into this class. The main result in the paper is that GMR tests control the experimentwise error, the probability of declaring at least one inactive effect active. Voss (1988) considered for example a method based on the smallest 50% of the mean squares ($f(x) = x^2$) in which a_i is a constant $1/m$ for the smallest $m(= 0.5n)$ unsigned contrasts and zero, otherwise.

Benski (1989)

Benski (1989) proposed using a modified Shapiro-Wilk test for normality (Shapiro and Francia (1972)) to test the presence of active effects coupled with an outlier test for identifying the particular effects that are active. The motivation for the Shapiro-Wilk test is a ratio of two estimates of variation, the squared estimated slope of the probability plot regression line and the standard deviation of the contrasts. The modified Shapiro-Wilk statistic W' is

$$W' = \Big(\sum_{i=1}^{k} z_{(i)k} c_{(i)}\Big)^2 / \Big(\sum_{i=1}^{k} z_{(i)k}^2 \sum_{i=1}^{k} (c_{(i)} - \bar{c})^2\Big), \tag{8}$$

where $\bar{c}$ is the average of the ordered contrasts $c_{(i)}$ and $z_{(i)k}$ are expected standard normal order statistics in a sample of size k. Normality is rejected for small values of W' which in this context corresponds to the contrasts all not having the same mean (i.e., some are non-zero). Since (8) can also be viewed as a correlation-type statistic (i.e., the mean of $z_{(i)k}$ is exactly zero), a large value (close to one) indicates a strong association between the normal distribution and the observed data. Consequently, small values of W' are taken to indicate the presence of at least one active effect. Note that the original Shapiro-Wilk test uses constants a_i based on best linear unbiased estimation rather than the $z_{(i)k}$ based on least-squares estimation presented here.

Once the Shapiro-Wilk test indicates the presence of active effects, Benski (1989) proposed using an outlier test to identify the active effects. The outlier test is based on a robust estimate of spread which uses the assumption of zero mean for the inert effects to arrive at the interval $(-2d_F, +2d_F)$, where d_F is the interquartile range, the difference between the first and third quartiles of the contrasts c_i. Those contrasts falling outside the interval are candidates for active effects. Benski (1989) proposed the following procedure: if the Shapiro-Wilk test is rejected, combine the p-values of both tests and declare the largest contrast active if the combined test is rejected. Then, drop the largest contrast and perform the same procedure on the remaining contrasts.

A comment about the first test in Benski's (1989) proposal is worthwhile. The Shapiro-Wilk test does not account for the arbitrariness of the factor level labels. Shapiro and Wilk (1965) noted this drawback in applying their test

statistic to data from a factorial experiment. Also, the test does not use the information that the mean of the inert contrasts is zero. This suggests using a half-normal version with the unsigned contrasts $|c_i|$ which will be presented in Section 3. Also note that the second test is almost the same as the outlier test used by Seheult and Tukey (1982).

Lenth (1989)

Lenth (1989) considered a robust estimator of the contrast standard error τ, which he termed the pseudo standard error estimate or PSE:

$$\text{PSE} = 1.5 \cdot \text{median}_{\{|c_i|<2.5s_0\}}|c_i|, \tag{9}$$

where

$$s_0 = 1.5 \cdot \text{median}|c_i|. \tag{10}$$

That is, PSE is a trimmed median which attempts to remove contrasts corresponding to active effects. Active effects are then identified using the margin of error ME=$t_{0.975;df}$PSE with degrees-of-freedom $df = k/3$ or the simultaneous margin of error SME=$t_{\gamma;d}$PSE, where $\gamma = (1 + 0.95^{1/k})/2$. Note that PSE is asymptotically normal (Dong 1993) and is consistent for τ when there are no active effects but overestimates τ, otherwise. The degrees-of-freedom $k/3$ come from an approximation of PSE^2 by a scaled χ^2 distribution. Using the PSE to standardize the contrasts gives statistics of similar form as in Daniel (1959) and Zahn (1975a):

$$|c|_{(i)}/\text{PSE}. \tag{11}$$

Bissell (1989)

Bissell (1989) proposed using Bartlett's (1937) test for variance homogeneity to identify the presence of active effects using the statistic

$$B = \ln((1/k)\sum M_i) - (1/k)\sum \ln(M_i), \tag{12}$$

where $\exp(B)$ is the ratio of the arithmetic mean of the mean squares to their geometric mean.

Bissell (1989) proposed using B sequentially for which the critical value at the ith stage is based on the remaining $k - i + 1$ effects being inert; the critical value is based on an appropriate F distribution.

Berk and Picard (1991)

Berk and Picard (1991) used the 60% smallest mean squares assuming that they correspond to inert effects to test the remaining larger mean squares with the statistic:

$$M_{(l)}/\sum_{i=1}^{m} M_{(i)}. \tag{13}$$

This is similar to Holms and Berrettoni (1969) except that m is fixed here rather than being determined by the contrasts. The critical values given in their Table 1 were computed under all inert effects and take account of the m smallest mean squares being the m smallest order statistics in a sample of size k. Berk and Picard (1991) commented that this formalizes Taguchi's (1987) approach of pooling the smallest mean squares by accounting for their true distribution. Voss (1988) considered the same method except that he based it on the 50% smallest mean squares.

Bissell (1992)

When there are no active effects, all the mean squares M_i have the same scaled χ^2 distribution, whose variance is a function of its mean. This relationship between the theoretical mean and variance provided the motivation for Cochran's (1954) dispersion tests which evaluates whether the relationship is supported by the data. Letting $\bar{M}$ and S_M^2 denote the sample mean and variance of the mean squares, respectively, then the test statistic is the coefficient of variation for the M_i's:

$$S_M/\bar{M}, \tag{14}$$

where k is the number of mean squares. The test rejects for large values with critical values for $S_M/\bar{M}$ being based on the approximation that $((k-1)/2)(S_M/\bar{M})^2 \sim \chi^2_{(k-1)}$, given in Table 12 of Bissell (1992) for $k = 2(1)31$.

Bissell (1992) suggested dropping several mean squares that are obviously active and then retesting the remaining effects. That is, the critical value for the test statistic is based on the remaining effects being inert from the corresponding sample size. Note that the χ^2 approximation does not account for the fact that the estimate $\bar{M}$ is used rather than the true mean.

Le and Zamar (1992)

Le and Zamar (1992) proposed using an outlier test based on the ratio of two estimates of scale, a non-robust estimate divided by a robust one. They suggested using two M-estimates S_1 and S_2 of τ which satisfy

$$(1/k)\sum_{i=1}^{k}\rho[(c_i - T)/S] = E(\rho(Z)), \tag{15}$$

where Z has a standard normal distribution, and whose ρ-functions are

$$\rho_1(x) = \begin{cases} x^2, & \text{if } |x| < a, \\ a^2, & \text{otherwise,} \end{cases} \tag{16}$$

and

$$\rho_2(x) = \rho_1 + \beta(x^4 - 6x^2). \tag{17}$$

Using the statistic

$$R_{\text{LZAM}} = S_2/S_1, \tag{18}$$

they proposed a sequential procedure by dropping the largest contrast and then recalculating (18) with the remaining contrasts. The critical values are based on all effects being inert for a sample size equal to the remaining number of contrasts. Note the similarity with the first part of Loh's (1992) proposal which also uses a ratio of a robust and non-robust estimates of scale.

A practical problem with ρ_2, however, is that it has two roots. To avoid this problem, another non-robust estimator could be used such as one based on

$$\rho_2^*(x) = x^2. \tag{19}$$

Juan and Pena (1992)

Juan and Pena (1992) proposed standardizing the contrasts by a different estimator for τ. It is similar to Lenth's (1989) PSE except that the calculation is iterative as follows: (a) Defining MAD_0 as the median of the k unsigned contrasts, recompute the median of those unsigned contrasts not exceeding $w\text{MAD}_0$ for some constant $w > 2$. Continue until the median stops changing and denote this by IMAD_0. (b) Then the estimator for τ is:

$$\hat{\tau}_{\text{IMAD}} = \text{IMAD}_0/a_w, \tag{20}$$

where a_w is a correction factor (See their Table 1 for a_w for a range of w.) Juan and Pena (1992) recommended w=3.5 and $a_w = 0.6578$ and showed that IMAD_0 has better MSE than PSE (11) when more than 25% of the effects are active. They also showed that the estimator based on the interquartile range d_F behaves poorly and that using the trimmed median is generally better than the trimmed mean when more than 20% of the effects are active.

Their procedure for identifying active effects can then be put in terms of the statistics:

$$|c|_{(i)}/\hat{\tau}_{\text{IMAD}}, \tag{21}$$

whose distribution is approximated by a standard normal distribution.

Loh (1992)

The motivation for Loh (1992) was to formally extend the graphical normal plot. Noting that the arbitrariness of labels yields different normal plots, Loh (1992) chose the set of contrasts with median closest to zero; in the case of ties, the one with largest correlation coefficient of the regression line on the normal probability plot is chosen. (This is related to the Shapiro-Wilk goodness-of-fit idea.) Like Benski (1989), it is a hybrid procedure. The initial test determines the presence of active effects by comparing the slope of the least-squares line

through all contrasts versus the slope of line through a set of smaller contrasts thought to be inert. The inert contrasts are those whose magnitude are less than twice d_F, the interquartile range (see Seheult and Tukey (1982) and Benski (1989)). The test is rejected for large ratio values with the outliers then becoming potential active effects. For identification, Loh (1992) proposed using the Scheffé prediction interval based on the fitted line to the inliers in the previous test; i.e., those outliers falling outside the prediction interval are identified as active.

Note that some computation is required in finding the set of contrasts used in the normal plot. Working with the unsigned contrasts eliminates all this computation, however; this suggests using a half-normal version which will be considered in Section 3.

Dong (1993)

Similar to Lenth (1989), Dong (1993) proposed an estimator for τ but based it on the trimmed mean of squared contrasts rather than the trimmed median of the unsigned contrasts: $s_{\mathrm{DONG}} = \sqrt{m^{-1}\sum_{\{|c_j|<2.5s_0\}} c_j^2}$, where m is the number of terms being summed and s_0 is defined earlier in (10). Dong (1993) showed that s_{DONG} has smaller MSE than PSE which provided his motivation for using it to standardize the contrasts as

$$|c|_{(i)}/s_{\mathrm{DONG}} \tag{22}$$

and suggested using $t_{\gamma,m}$ as the critical value for suitable choice of γ. Dong (1993) also proposed iteratively calculating s_{DONG} until it stops changing when there are a large number of active effects.

Schneider, Kasperksi and Weissfeld (1993)

Schneider, Kasperksi and Weissfeld (1993) proposed standardizing the contrasts by an estimator of τ given in Wilk, Gnanadesikan and Freeny (1963); by treating the m smallest unsigned contrasts all thought to be inert as a Type II right-censored sample, τ can be estimated using the maximum likelihood estimator (MLE) $\hat{\tau}_{\mathrm{CEN}}$. The MLE does not have a closed form, however. See details in Schneider et al. (1993). Their motivation for treating the contrasts as a censored sample was to reduce the bias and suggests the following standardized contrasts:

$$|c|_{(i)}/\hat{\tau}_{\mathrm{CEN}}. \tag{23}$$

Schneider et al. (1993) use the asymptotic normality of (23) to calculate approximate critical values.

Venter and Steel (1996)

Venter and Steel (1996) proposed using a procedure which first tests whether all effects are inert and, if rejected, identifies the active contrasts causing rejection.

Their procedure uses successive ratios V_i, where

$$V_i = |c|_{(i+1)}|/\sqrt{(1/i)\sum_{j=1}^{i}|c|_{(j)}{}^2}, \tag{24}$$

whose corresponding p value is $P_i = 1 - F_i(V_i)$, where $F_i(x) = Prob(V_i \leq X|$ all k effects are inert). Assuming effect sparsity, i.e., there are at least l inert effects,

$$S_l = \min(P_i : l \leq i \leq k-1) \tag{25}$$

is used to test that all effects are inert. If $S_l \leq s_l(\alpha)$, the test is rejected and the active contrasts causing rejection test are identified by the first index $\hat{q} \geq l$ such that $P_{\hat{q}} \leq s_l(\alpha)$. $s_l(\alpha)$ is the αth quantile of the S_l distribution which is given in their Table 2. Note the similarity of the strategies of Venter and Steel (1996) and Holms and Berrettoni (1968).

3. Modifications and New Proposals

Some modifications of existing methods as well as new proposals will be considered next.

Modified Loh (1992)

As suggested by Loh (1992), a formalization of the half-normal plot of the unsigned contrasts can be done as follows: (a) the inliers are those not exceeding four times the median of the unsigned contrasts; (b) fit the least-squares line through origin of all ordered unsigned contrasts against their respective expected standard half-normal order statistics to obtain a slope estimate $\hat{\beta}_1$; (c) fit the least-squares line through origin of the ordered set of inliers defined in (a) against their respective expected standard half-normal order statistics to obtain a slope estimate $\hat{\beta}_2$; (d) the test for presence of active effects is based on $R = \hat{\beta}_1/\hat{\beta}_2$ which rejects for large values of R; (e) identify the active effects corresponding to those outliers exceeding the prediction interval based on fitted line to the inliers in (c) above; i.e., $||c|_{(l)} - \hat{\beta}_2|z|_{(l)k}| > S_2(k'F_{k',m-1;\gamma})^{1/2}(1+w)^{1/2}$ where m is the number of inliers, $k' =$ ceiling$[k/4]$, S_2 is the root mean squared error of the fitted line in (c), and $w = |z|^2_{(l)k}/\sum_{i=1}^{m}|z|^2_{(i)k}$.

Modified Schneider et al. (1993) and Lenth (1989)

Schneider et al. (1993) and Lenth (1989) estimate τ based on censoring and trimming. This could be done sequentially by dropping the largest contrast and applying the procedures on the remaining contrasts whose sample size is one less. The critical values would then be calculated under the reduced sample size at each stage.

Probability Plot Correlation Coefficient

As a measure of linearity of a probability plot, Filliben (1975) proposed calculating the correlation coefficient between the ordered contrasts $c_{(i)}$ and the median standard normal order statistics $\tilde{z}_{(i)k}$.

$$R_{\mathrm{CORR}} = \sum_{i=1}^{k}(\tilde{z}_{(i)k} - \bar{\tilde{z}})(c_{(i)} - \bar{c})/\Big(\sqrt{\sum_{i=1}^{k}(\tilde{z}_{(i)k} - \bar{\tilde{z}})^2}\sqrt{\sum_{i=1}^{k}(c_{(i)} - \bar{c})^2}\Big). \qquad (26)$$

Note the similarity with the modified Shapiro-Wilk statistic W' in (8) except that medians are used instead of means. Again because of the arbitrariness of the labels, we will consider a half-normal version which uses unsigned contrasts $|c_i|$ and expected standard half-normal order statistics $|z|_{(i)k}$ (instead of medians) in (26) above. Small values of R_{CORR} suggest the presence of active effects. The procedure could be used sequentially with critical values being calculated for the reduced sample size at each stage.

Half-Normal Shapiro-Wilk Test

While Shapiro-Wilk (1965) suggested a half-normal version, it has apparently not been discussed further in the literature. In the present context, it is natural to consider this version since working with the unsigned contrasts $|c_i|$ removes the arbitrariness of the labels. Using the means, variances and covariances of the standard half-normal order statistics tabulated by Govindarajulu and Eisenstat (1965), the Best Linear Unbiased Estimator (BLUE) of τ based on the m smallest order statistics is given by (see Balakrishnan and Cohen (1991), p. 74)

$$\hat{\tau}_{\mathrm{BLUE}} = \underset{\sim}{\mu}^T \Sigma^{-1} |\underset{\sim}{c}|_{(\ )} / (\underset{\sim}{\mu}^T \Sigma^{-1} \underset{\sim}{\mu}), \qquad (27)$$

where $|\underset{\sim}{c}|_{(\ )}$ denotes the vector of m smallest $|c_i|$, $\underset{\sim}{\mu}$ is the vector of the means of the m smallest standard half-normal order statistics in a sample of size k and $\sum$ is the variance-covariance matrix of these order statistics. (See Tables A1 and A2 in the Appendix for the coefficients used to compute (27) for $n = 8$ and $n = 16$, respectively.) Since the MLE of τ based on the $|c_i|$ values is

$$\hat{\tau}_{\mathrm{MLE}} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}|c_i|^2},$$

we consider a Shapiro-Wilk type goodness-of-fit test given by

$$\mathrm{HSW} = \hat{\tau}_{\mathrm{BLUE}}/\hat{\tau}_{\mathrm{MLE}}, \qquad (28)$$

which suggests the presence of active effects for small values of HSW. This statistic can be used sequentially by removing the largest unsigned contrast and so

forth with critical values being calculated for the reduced sample size at each stage. Analogous to the Shapiro-Wilk test, the critical region is taken to be small values of HSW which has been confirmed by empirical analysis.

Dispersion Test

Since the $|c_i|$ have a half-normal distribution (under all inert effects), the ordered $|c|_{(i)}$ on average should be close to $\tau|z|_{(i)k}$. Consequently, we propose a *dispersion test* procedure based on the m smallest $|c_i|$ values, using the statistic

$$D_m = \frac{1}{m}\sum_{i=1}^{m}\Big(\frac{|c|_{(i)}}{PSE\ |z|_{(i)k}} - 1\Big)^2. \tag{29}$$

Note that since PSE is a "robust" estimator of τ, a significant departure of $|c|_{(i)}$/PSE from its expected value $|z|_{(i)k}$ (under all inert effects) suggests an active contrast so that the test rejects for large values of D_m in (29). This statistic can also be used sequentially with critical values being calculated for the reduced sample size at each stage.

4. A Comparison of the Methods

First, similarities in the form of the methods will be presented in Section 4.1. Then a comparison of their performance based on a simulation study will be discussed in Section 4.2.

4.1. An initial comparison

Grouping the methods into the following broad categories is helpful:

Directed vs. Composite Directed methods test the individual effects directly. DAN59, DONG93, JP92, LEN89, SKW93 and ZAHN75 (also the modified versions MDONG, MLEN, MSKW) standardize the contrasts by various estimates of σ. BM86 and JTUK87 use individual posterior probabilities and ratio-to-scale statistics, respectively. HB69, BP91 and VS96 use mean squares. The second parts of BEN89 and HLOH92 test are also directed; BEN89's second part is the same as STUK82 and that of HLOH92 is related. The composite methods test all the effects as a group. These include BI89, BI92, CORR, DISP, HSW, MLZ92, as well as the first parts of BEN89 and HLOH92.

Sequential BI89, BI92, DAN59, HB69, JP92, JTUK87, MLZ92, STUK82, and ZAHN75 as proposed are sequential meaning that some computation is done at each stage with the remaining contrasts. For all the sequential methods except DAN59, the critical values are based on all inert effects for the current sample size at a given stage; DAN59 bases its critical values on all $k = n - 1$ effects

being inert. The other methods not listed above are not sequential but a suitable version could be developed.

Hybrid BEN89 and HLOH92 are hybrids of two methods.

4.2. A simulation study

Limited studies comparing only some of the existing methods listed above have been done: Zahn (1975b), Voss (1988), Berk and Picard (1991), Loh (1992), Dong (1993), Haaland and O'Connell (1995), Benski (1995) and Benski and Cabau (1995). Because the off-the-shelf performance of these methods is not the same when all effects are inert, it is difficult to compare the power of the various methods directly. Table 1 gives the off-the-shelf performance of the existing methods presented in Section 2 when all effects are inert. Note that the half-normal version of Loh (1992) given in Section 3 is used rather than the original full normal version. Based on 10,000 simulations for a 16 run experiment ($k = 15$), Table 1 gives the observed proportion of simulations when zero to eight effects were declared active under all effects being inert. Note that no two procedures have exactly the same performance.

Table 1. Off-the-shelf performance of existing methods p_i = observed proportion of simulations detecting i effects under all inert effects for 16 run design (* indicates ≥ 8 declared effects)

	number of declared effects										
method	0	1	2	3	4	5	6	7	8	IER	EER
BEN89	.975	.020	.003	.002	.000	.000	.000	.000		.0022	.025
BI89	.948	.023	.007	.004	.002	.003	.002	.002	.009*	.0139	.052
BI92	.834	.118	.032	.011	.004	.001	.000			.0157	.166
BM86	.748	.176	.044	.016	.007	.004	.003	.002	.000	.0262	.252
BP91	.555	.259	.119	.050	.017	.004	.000			.0492	.445
DAN59	.598	.193	.093	.050	.065					.0527	.402
DONG93	.569	.302	.085	.029	.011	.004	.001	.000		.0418	.431
HB69	.629	.157	.067	.045	.033	.029	.042			.0634	.371
HLOH92	.951	.017	.018	.010	.004	.001	.000			.0070	.049
LEN89	.755	.144	.054	.024	.013	.007	.003	.001		.0290	.245
JP92	.799	.104	.039	.021	.014	.010	.006	.004	.003*	.0294	.201
JTUK87	.950	.034	.010	.003	.002	.001	.001	.000	.000	.0054	.050
MLZ92	.953	.023	.007	.007	.003	.003	.002	.001	.001	.0074	.047
SKW93	.590	.254	.105	.038	.011	.002	.000			.0421	.410
STUK82	.742	.129	.054	.026	.017	.012	.008	.005	.005*	.0387	.258
VS92	.900	.083	.015	.001	.000	.000				.0079	.100
ZAHN75	.618	.190	.089	.048	.055					.0487	.382

Two summary measures which will be useful for reference are the experimentwise error rate (EER) and the individual error rate (IER). Let p_i denote the proportion of simulations for i inert effects declared active. Then EER is proportion of the simulations when one or more effects is declared active, $1 - p_0$. The IER is the average proportion of inactive effects declared active, $\sum(i/(n-1))p_i$. This definition of IER when all effects are inactive can be extended to the case when some effects are active by suitably changing $n-1$ to the number of inactive effects. Note that the EER and IER given in Table 1 vary across the different methods.

The different off-the-shelf performance of these methods depend in part on how they were designed which often involve the IER or EER criteria. DAN59 (critical values from Zahn 1975a) and ZAHN75 attempt to control IER at 0.05. Note that DAN59 and ZAHN75 as used here can detect at most four effects. HB69 was started by pooling the nine smallest effects ($m = 9$) and used α_{pool}= 0.25 and α_{final}=0.05 so that IER at 0.05 is implied. LEN89, SKW93 and DONG93 as reported here attempt to control IER at 0.05; differences for these tests arise from approximate distributions used in calculating the critical values. Also, an attempt to control EER can be done using a suitable choice of IER based on simultaneously testing k contrasts per experiment; this is the basis for JP92 which attempts to control EER at 0.05 (but still turns out to be as large as 0.201). BM86 uses $(\alpha_{active}, K) = (0.2, 10)$ and a marginal posterior probability threshold of 0.5. There were no parameters to set for STUK82. BEN89 used 0.05 levels for the normality test (for presence of active effects) and the pooled normality-outlier test (for identification of active effects); thus, the initial test attempts to control the EER at 0.05. BP91 controls IER exactly at 0.05. HLOH92 used a 0.05 level test for the presence of active effects and a 95% simultaneous prediction interval for identifying the active effects; consequently, EER is controlled at 0.05. JTUK87 attempts to control IER at 0.05 (values for 11-14 are not given in Johnson and Tukey (1987) and were simulated based on 10,000 samples). BI89, BI92 and MLZ92 (MLZ92 uses (19) instead of (17).) as reported here control EER at 0.05, whereas VS96 controls EER at 0.10.

The challenge then is to compare these methods on as equal a basis as possible without destroying the essence of the methods. Comparing them on an exactly equal basis is not possible because of the different forms of the methods as listed in Section 4.1. For example, the sequential methods control EER while the non-sequential methods control IER. STUK82, BEN89 and HLOH92 are hard to change and best evaluated as originally proposed. An earlier version of this paper (Balakrishnan and Hamada 1994) compared directed sequential versions of all the methods (except BEN89 and LOH92) which were controlled to have

the same EER and IER; that is, the absolute contrasts were standardized by a suitable measure. One of the referees asked whether this standardization had destroyed the essence of some of the methods. Consequently, the simulation study was redone to use the methods as originally proposed. In order to compare the power of these methods fairly, each has been adjusted so that IER for all inert effects is controlled at 0.044; the critical values are based on a simulation using 100,000 samples. Note that STUK82's IER of 0.038 is close to the others with 0.044. To compare BEN89 and HLOH92, BP91 was also studied at an IER of 0.007 (BEN89 and HLOH92's IERs are 0.002 and 0.007, respectively). See Figure 1 which gives the EER and IER of the methods and shows that IER is near 0.044 for the non-hybrid methods and near 0.007 for the hybrid methods. Note that each method has a pair of values. The left value is based on normal errors. The right value is based on errors with a standardized Student t distribution with nine degrees of freedom, where these errors were standardized to have variance one. The study with Student t errors was undertaken to address a question raised by one of the associate editors regarding the robustness of the methods to nonnormality; this will be discussed later.

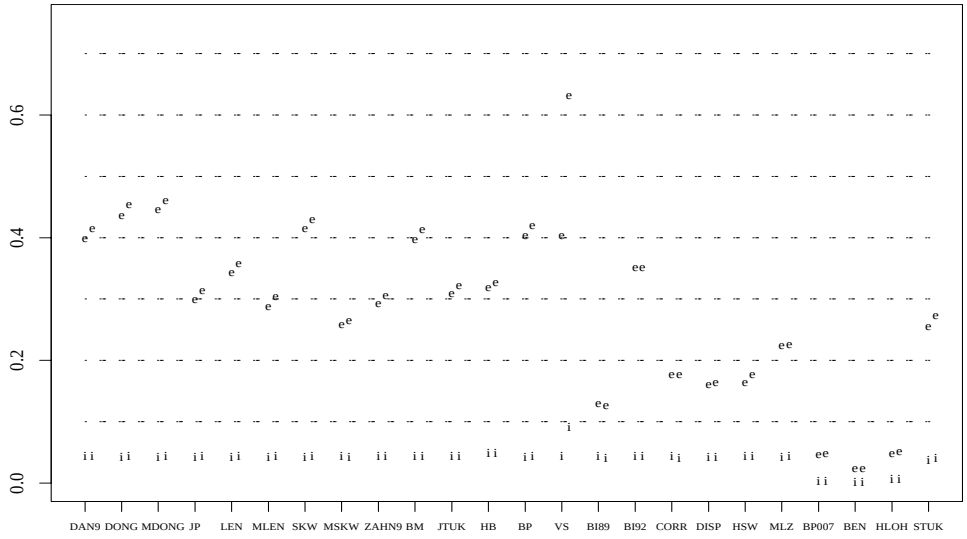

Figure 1. EER, IER for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ no active effects (e =EER, i =IER)

Some details for the particular versions of the methods used in the study follows. HB69 uses ($m = 7$, $\alpha_{pool} = 0.50$, $\alpha_{final} = 0.01$), the version given in the original paper with an IER closest to 0.044 (0.049). BP91, SKW91 and

VS96 use $m = 9$ so that the six largest contrasts are tested. The sequential methods DAN59, MLEN, MSKW, ZAHN75, JTUK87, BI89, BI92 and MLZ92 test the eight largest contrasts; MLEN and MSKW are the modified versions of LEN89 and SKW93 which account for k', the current number of contrasts being considered at a given stage. The estimators of contrast standard deviation for ZAHN75 and DAN59 are based on the nine smallest contrasts instead of 11 so that up to six active effects could be detected. MDONG refers to the procedure which uses the iterative estimate of τ based on (22) proposed by Dong (1993).

Since most of the methods have been adjusted so as to have the same IER under all inert effects, their power can be investigated under various scenarios. This was also done by simulation based on 10,000 samples. For $n = 16$ runs, one, two, four and six active effects all having the same magnitude from 0.5σ to 4σ were studied (.5(.5)3,4 σ). Recall that σ denotes the process or error standard deviation not the contrast standard error. Active effects with the same magnitude were used because they provide bounds on the performance of when the effects have different magnitudes. Note that the value at 0.0σ is the method's size which is 0.044 for all the methods except BEN89, HLOH92 and STUK82. Figures 2-5 display the power (or average proportion of active effects that were declared active).

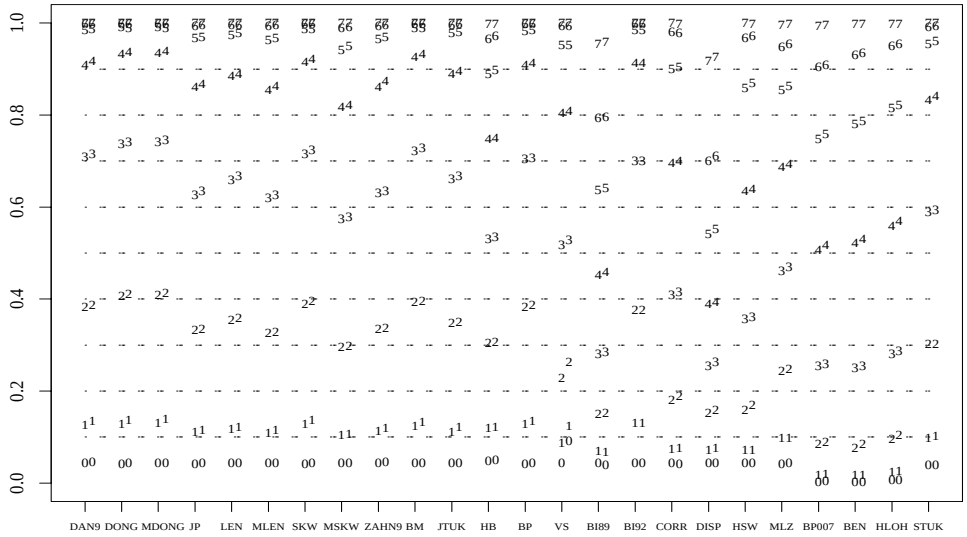

Figure 2. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ one active effect $= 0, .5(.5)3, 4\ \sigma$ (labels 0-7)

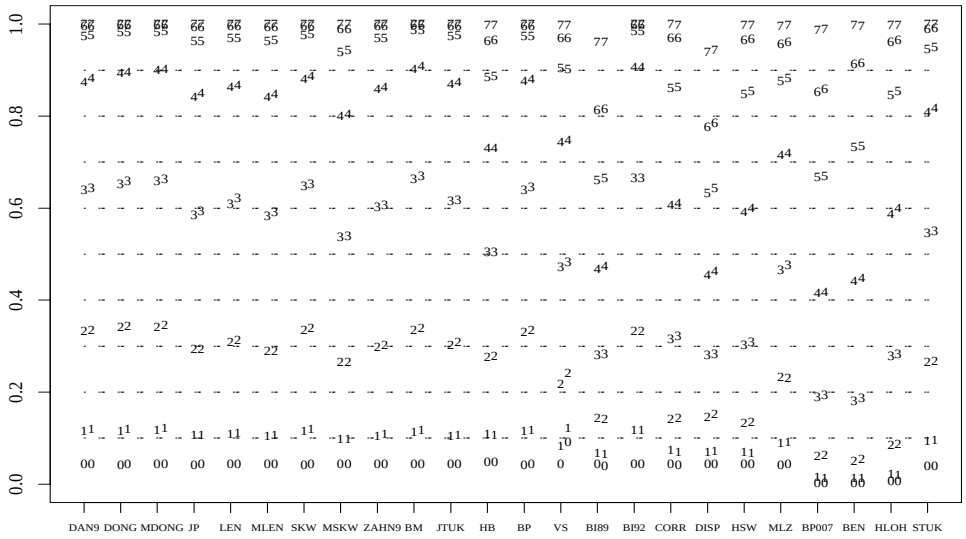

Figure 3. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ two active effects $= 0, .5(.5)3, 4\ \sigma$ (labels 0-7)

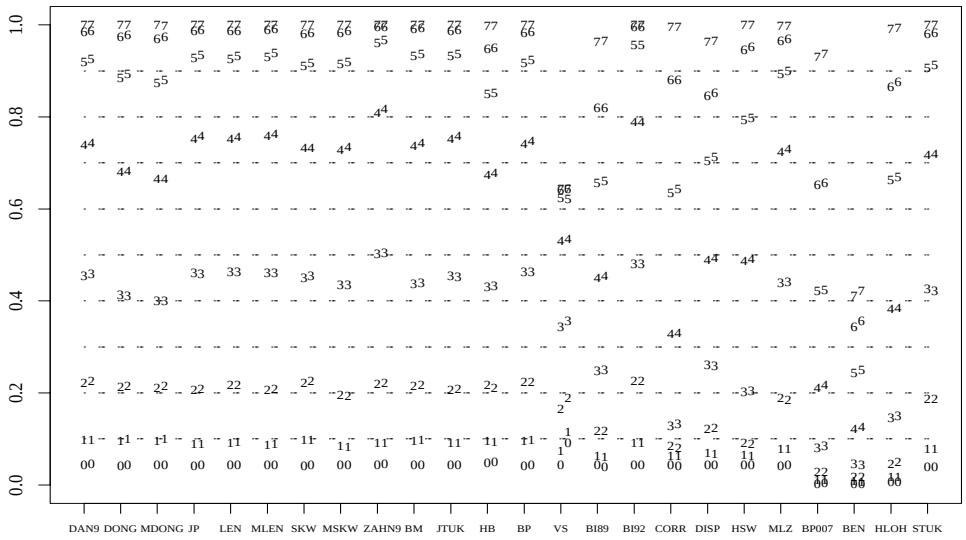

Figure 4. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ four active effects $= 0, .5(.5)3, 4\ \sigma$ (labels 0-7)

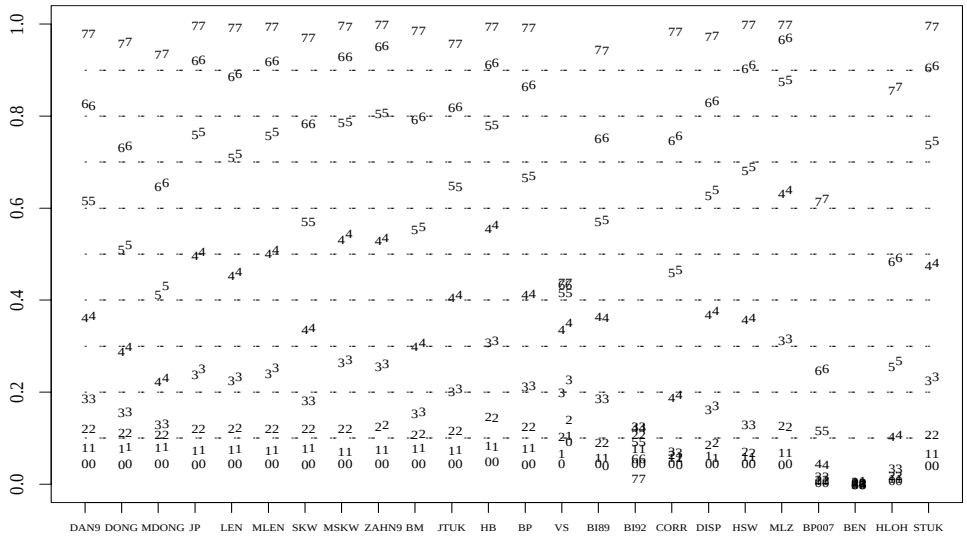

Figure 5. Power for $N(0,1)$ and std. $t(9)$ errors (1st, 2nd of pair) and $n = 16$ six active effects $= 0, .5(.5)3, 4\ \sigma$ (labels 0-7)

Some conclusions from the simulation study based on normal errors follow in which #AE denotes the number of active effects:

- There is little difference between the methods for small size effects, say 0.5σ, which exibit little power. Also, there is not much of a difference for large size effects, say 4σ. There are marked differences between the methods for intermediate size effects ($\sigma - 3\sigma$) which become more pronounced for larger size effects in this range as the #AE increases; there is less of a difference for smaller size effects as the #AE increases.
- The power decreases as the number of active effects increases.
- Except for BI92 (see the six active effect case), the power increases as the size of the active effects increases. Note that the effects need to be rather large relative to the process standard deviation σ. For example, the power is around 0.7 for a single 1.5σ effect.
- The directed methods which focus on the current largest unsigned contrast tend to perform better than the composite methods, BI89, BI92, MLZAM, CORR, DISP and HSW. HSW is a goodness-of-fit procedure which tests for any violation of half-normality and is not directed specifically for detecting extreme values; this explains why its power is not as high as those which are so directed. MLZ92 is an exception which performs surprisingly well, especially for large #AE, however. CORR is clearly the worst of all the composite methods.

- BI92, a composite procedure appears promising say for up to four active effects but then its performance seriously degrades for six active effects. This can be explained since the variance of the mean squares will tend to decrease when there are too many active effects (i.e., the roles of the inert and active contrasts are switched) while their mean increases resulting in small values for (14). This is clearly an undesirable property.
- Many of the directed methods only differ in the estimator used for τ. Various proposals were motivated by better MSE properties of the estimators. For example, S_{ZAHN} outperforms $|c|_{(11)15}$ (Zahn (1975b)). Juan and Pena (1992) showed that IMAD_0 performed better than Lenth's (1989) PSE and an estimator based on d_F performed much worse than both IMAD_0 and PSE. SKW93 was motivated similarly with censoring being used to reduce the bias of the estimator. DONG93 used trimmed means instead of the trimmed medians used by PSE because of improved efficiency. Nair (1984) also compares different estimators based on the full normal probability plot. Yet, the gains in estimator performance appear to have little impact on the test performance. Rather, the #AE seems to affect two groups of the methods differently. DAN59, DONG94, MDONG and SKW93 do better than JP92, LEN89, MLEN, MSKW and ZAHN75 for small #AE whereas the latter group perform better for large #AE. Overall, DAN59 and SKW93 perform the best in the first group. ZAHN75 performs the best in the second group although there is not much of a difference between these methods.
- Among the other directed methods (BM86, JTUK87, HB69, BP91, VS96), BP91 performs the best, although BM86 is quite competitive for small #AE; BM86 performs poorly for six active effects, but recall that α_{prior} was set at 0.2, i.e., it was designed for three active effects.
- The modified procedures MLEN and MSKW provide little if any improvement over LEN89 and SKW92, except for large #AE.
- MDONG has almost the same power as DONG93 for small #AE, and actually performs worse as #AE increases. Thus, there is no real benefit offered by the iteration in estimating τ.
- STUK82 is quite competitive with the best of the other methods for #AE up to four and is slightly worse for six active effects. Recall that it is at a slight disadvantage since its IER is 0.038 as compared with 0.044 for the other methods.
- Among the two hybrid methods, HLOH92 performs better than BEN89 although BEN89's IER and EER is smaller. BEN89's performance degrades dramatically for large #AE. HLOH92 also outperforms BP91 which was adjusted to have the same IER of 0.007. Recall that BP91 is one of the best methods.

Besides power, the IER or average proportion of inactive effects that are declared active under the various scenarios needs to be studied. Such plots of IER are not shown here to conserve space; a summary will be given, instead. Recall that under all inert effects, the IER is 0.044. The plots show that the methods for the most part are conservative, i.e., IER is below 0.044 and decreases as #AE increases. The IER depends on the magnitude of the active effects and for many of the methods is a nonmonotonic function of active effect magnitude.

Robustness of Methods to Nonnormality

As mentioned previously, the methods were studied using standardized Student t with nine degrees of freedom errors to address the robustness of methods to nonnormality. Here, the error distribution is flatter than the normal distribution. Figure 6 displays contrasts for $n = 16$ based on this error distribution and shows that the contrasts are nearly normal. Recall that the contrasts are linear combinations of the observations, so that the Central Limit Theorem effect explains why the contrasts are more normal-like than the individual observations. Consequently, it is not surprising that the majority of the methods have nearly the same performance with these flatter distributed errors. (See the right values of the pairs in Figures 1-5.) The IER and consequently the power of the methods are slightly higher which can also be explained by the flatter error distribution. Note that VS96 is effected the most by the flatter error distribution with its size approximately doubling.

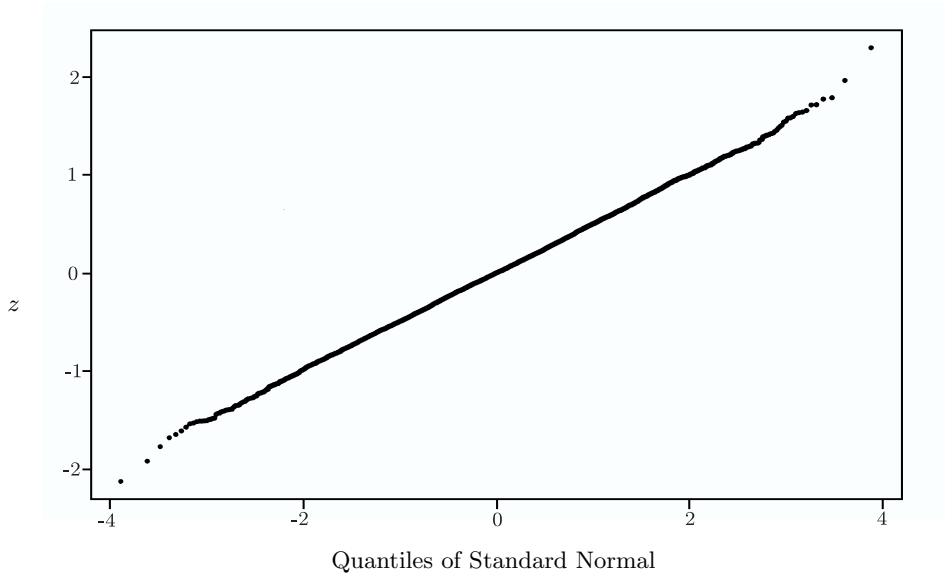

Figure 6. Normal probability plot of contrasts for $n = 16$ and std. $t(9)$ error

5. Discussion

Recently, several papers have proposed methods but have only compared their performance with some existing methods using some data sets. Based on this paper, some recommendations for evaluating new procedures are:

- A simulation study is needed to evaluate the performance of the new method with existing ones. Calibration of the new procedure needs to be done in order to provide a fair comparison. At a minimum, it should report the p_i values under all inert effects.
- In addition to examining the power of the proposed procedure, its IER behavior should also be studied.
- One should check to see if the method is exploiting the following properties of the contrasts: (1) The contrasts have equal variances and are normally distributed. (2) Contrasts for inert effects have zero means while those for estimating active effects have non-zero means.
- The method should not depend on the arbitrariness of the factor level labels. For example, methods that work with unsigned or squared contrasts avoids this problem.

In the course of this research, several issues and possibility of connections with other areas in statistics arose which warrant more study.

- What are desirable EER, IER and p_i when all effects are inert? For example, using an IER of 0.044 for $n = 16$, some methods had an EER of 0.40, which some might consider large.
- Should other measures of performance be used such as an overall performance measure that accounts for both a procedure's ability to detect active effects as well as its tendency to identify inactive effects as active? See Benski (1995) and Benski and Cabau (1995) who make one proposal.
- Should other non-inert scenarios be considered such as different sized active effects? If so, what would be appropriate choices for the sizes of the active effects? See Holms and Berrettoni (1969), Haaland and O'Connell (1995), Benski (1995) and Benski and Cabau (1995) for some different scenarios.
- Are there non-sequential procedures which have better performance or are sequential directed tests preferable?
- Can gains be made by combining methods, i.e., hybrid methods? The simulation study showed that the half normal version of Loh (1992) is promising.
- Can information on how many active effects there are likely to be present in the experiment be exploited? The simulation results suggest an affirmative answer, but which methods are less sensitive to such a specification?
- How robust are the methods to other nonnormal distributions? In such a situation, would a nonparametric method be preferable? For example, Loughin and Noble (1997) propose a permutation test procedure.

- There are connections with other areas of statistics. For example, the work of Le and Zamar (1992) drew on the robust statistics literature. Seheult and Tukey (1982) and Benski (1989) viewed the active effects as outliers which has an extensive literature (Barnett and Lewis 1994). The ranking and selection (Gupta and Panchapakesan 1979) and multiple comparison (Hochberg and Tamhane 1987) literatures are also likely to be relevant. It will be interesting to explore how these different areas may help in suggesting new and possibly optimal tests and alternative ways to evaluate such methods.

6. Specific Recommendations

To conclude, we briefly summarize the results of our simulation study in terms of specific recommendations for the practitioner. The recommendations are:

- For up to six active effects, overall DAN59, SKW93, ZAHN75, BP91 and HLOH92 performed well, although others are competitive if one has a good idea about how many active effects there are. For example, for eight active effects, the versions of DAN59 and ZAHN75 used here would not be expected to do well since they assumed there would be no more than six active effects.
- The power for BI92 seriously degrades when there are many active effects so that this method is not recommended. For the same reasons, BEN89 is also not recommended.
- Especially for $n = 8$, the substantial variability exhibited in the probability plots when all effects are inert makes it difficult to both identify the active effects and to not choose the inert effects. Objective methods, which directly account for this variability, are therefore preferable. Nevertheless, Balakrishnan and Hamada (1994) showed that for such a small run size, the active effects need to be large relative to the process standard deviation for any hope of detecting them. Consequently, a larger run size ($n = 16$) is recommended.

Acknowledgements

We thank Fred Hulting for kindly sending his FORTRAN program PSTPRB as described in Stephenson, Hulting and Moore (1989). We also thank Wei-Yin Loh, Dan Meyer, Perry Haaland, Clif Young, Jock MacKay, two associate editors and three referees whose insightful comments on earlier versions helped to improve this paper. N. Balakrishnan's research was carried out while he was on sabbatical leave at the University of Waterloo and was supported by the Natural Sciences and Engineering Research Council of Canada. M. Hamada's research was supported by General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada.

Appendix

Tables used to implement HSW (28) are given in this appendix.

Half-Normal Shapiro-Wilk Test

The statistic $\hat{\tau}_{\mathrm{BLUE}}$ (27) can be written as a linear combination of the $k'(m = k')$ order statistics whose coefficients are given in Table A1 for $n = 8$ and Table A2 for $n = 16$.

Table A1. Half-normal Shapiro-Wilk test coefficients for $n = 8$ (order i corresponds to $|c|_{(i)k'}$)

	k'			
order	4	5	6	7
1	.262082	.215692	.183441	.159674
2	.553365	.447641	.376949	.326047
3	.911363	.711950	.589027	.504204
4	1.464728	1.044305	.834874	.702123
5	.0	1.569834	1.149021	.934437
6	.0	.0	1.653996	1.234854
7	.0	.0	.0	1.723853

Table A2. Half-normal Shapiro-Wilk test coefficients for $n = 16$ (order i corresponds to $|c|_{(i)k'}$)

	k'											
order	4	5	6	7	8	9	10	11	12	13	14	15
1	.0939	.0636	.0461	.0353	.0279	.0224	.0186	.0156	.0137	.0116	.0100	.0089
2	.1599	.1048	.0747	.0559	.0437	.0354	.0289	.0246	.0205	.0180	.0158	.0139
3	.2496	.1557	.1079	.0798	.0615	.0491	.0402	.0331	.0282	.0240	.0214	.0184
4	.4502	.2242	.1485	.1073	.0816	.0643	.0522	.0434	.0367	.0312	.0265	.0237
5	.0	.3786	.2031	.1406	.1047	.0817	.0657	.0542	.0453	.0390	.0338	.0292
6	.0	.0	.3279	.1857	.1330	.1015	.0805	.0658	.0549	.0466	.0398	.0348
7	.0	.0	.0	.2899	.1711	.1258	.0980	.0790	.0656	.0548	.0474	.0408
8	.0	.0	.0	.0	.2603	.1587	.1192	.0944	.0770	.0648	.0548	.0476
9	.0	.0	.0	.0	.0	.2366	.1481	.1132	.0908	.0750	.0634	.0543
10	.0	.0	.0	.0	.0	.0	.2170	.1388	.1077	.0876	.0733	.0625
11	.0	.0	.0	.0	.0	.0	.0	.2006	.1307	.1027	.0842	.0709
12	.0	.0	.0	.0	.0	.0	.0	.0	.1867	.1236	.0982	.0814
13	.0	.0	.0	.0	.0	.0	.0	.0	.0	.1746	.1173	.0940
14	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.1641	.1115
15	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.1549

References

Balakrishnan, N. and Cohen, Jr., A. C. (1991). *Order Statistics and Inference: Estimation Methods*. San Diego: Academic Press.

Balakrishnan, N. and Hamada, M. (1994). Analyzing unreplicated factorial experiments: A review with some new proposals. University of Waterloo Institute for Improvement in Quality and Productivity Research Report.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. Ser. A* **160**, 268-282.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition. Chichester, UK: John Wiley and Sons, Inc.

Benski, H. C. (1989). Use of a normality test to identify significant effects in factorial designs. *J. Quality Technology* **21**, 174-178.

Benski, C. (1995). Strategies for simulating active and noise effects in unreplicated experimental designs. 1995 *Proceedings of the Section on Quality and Productivity*, Alexandria, VA: American Statistical Association, 88-92.

Benski, C. and Cabau, E. (1995). Unreplicated experimental designs in reliability growth programs. *IEEE Trans. Reliability* **44**, 199-205.

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

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

Bissell, A. F. (1989). Interpreting mean squares in saturated fractional designs. *J. Appl. Statist.* **16**, 7-18.

Bissell, A. F. (1992). Mean squares in saturated fractional designs revisited. *J. Appl. Statist.* **19**, 351-366.

Box, G. (1988). Signal-to-noise ratios, performance criteria and transformations. *Technometrics* **30**, 1-17.

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

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ^2 Test. *Biometrics* **10**, 417-451.

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

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

Daniel, C. (1983). Half-normal plots. In *Encyclopedia of Statistical Sciences*, Volume 3 (Edited by S. Kotz and N. L. Johnson), John Wiley, New York.

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

Dong, F. (1993). Asymptotic properties of quantiles for truncated and contaminated data. *Comm. Statist. Theory Methods* **22**, 3255-3261.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics* **17**, 111-117.

Govindarajulu, Z. and Eisenstat, S. (1965). Best estimates of location and scale parameters of a Chi (1 d.f.) Distribution, Using Ordered Observations. *Rep. Statist. Appl. Res.* JUSE **12**, 149-164.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple decision procedures: theory and methodology of selecting and ranking populations.* John Wiley, New York.

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

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures.* John Wiley, New York.

Holms, A. G. and Berrettoni, J. N. (1969). Chain-pooling ANOVA for two-level factorial replication-free experiments. *Technometrics* **11**, 725-746.

Hurley, P. D. (1995). The conservative nature of the effect sparsity assumption for saturated fractional factorial experiments. *Quality Engineering* **7**, 657-671.

Johnson, E. G. and Tukey, J. W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In *Design, Data and Analysis* (Edited by C. L. Mallows), John Wiley, New York.

Juan, J. and Pena, D. (1992). A simple method to identify significant effects in unreplicated two-level factorial designs. *Comm. in Statist. Theory and Methods* **21**, 1383-1403.

Le, N. D. and Zamar, R. H. (1992). A global test for effects in 2^k factorial design without replicates. *J. Statist. Comput. Simulation* **41**, 41-54.

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

Loughin, T. M. and Noble, W. (1997). A permutation test for effects in an unreplicated factorial design. *Technometrics* **39**, 180-190.

Loh, W. Y. (1992). Identification of active contrasts in unreplicated factorial experiments. *Comput. Statist and Data Anal.* **14**, 135-148.

Nair, V. N. (1984). On the behavior of some estimators from probability plots. *J. Amer. Statist. Assoc.* **79**, 823-831.

Schneider, H., Kasperski, W. J. and Weissfeld, L. (1993). Finding significant effects for unreplicated fractional factorials using the n smallest contrasts. *J. Quality Technology* **25**, 18-27.

Seheult, A. and Tukey, J. W. (1982). Some resistant procedures for analyzing 2^n factorial experiments. *Utilitas Math.* **21B**, 57-98.

Shapiro, S. S. and Francia, R. S. (1972). Approximate analysis of variance test for normality. *J. Amer. Statist. Assoc.* **67**, 215-216.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (Complete Samples). *Biometrika* **52**, 591-611.

Stephenson, W. R., Hulting, F. L. and Moore, K. (1989). Posterior probabilities for identifying active effects in unreplicated experiments. *J. Quality Technology* **21**, 202-212.

Taguchi, G. (1987). *System of Experimental Design*. White Plains, NY: Unipub/Kraus International Publications.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Reading, MA: Addison-Wesley Publishing Company.

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

Voss, D. T. (1988). Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error. *Comm. Statist. Theory Methods* **17**, 3345-3359.

Wilk, M. B., Gnanadesikan, R. and Freeny, A. E. (1963). Estimation of error variance from smallest ordered contrasts. *J. Amer. Statist. Assoc.* **58**, 152-160.

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

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

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

E-mail: mshamada@m.imap.itd.umich.edu

Department of Mathematics and Statistics, McMaster University Hamilton, Ontario, Canada L8S 4K1

E-mail: bala@mcmail.cis.mcmaster.ca

(Received September 1994; accepted September 1997)

COMMENT

Claudio Benski

Schneider Electric

Hamada and Balakrishnan must be commended for have written a very extensive review concerning a wide variety of methods used to identify active effects in unreplicated factorial experiments. Economical and technical reasons have contributed considerable appeal to the ever-increasing use of unreplicated experimental designs in industrial settings. It is therefore important to determine what kind of approach should an experimenter adopt to assess the statistical significance of the considered factors in the absence of an independent noise estimate. The consequences of a mistake on this decision-making process can be enormous and industry would certainly be very interested in a clear answer in this area. Several researchers have previously tested some of the available methods and have reported them in the literature, for example, Haaland and O'Connell (1995) and Benski and Cabau (1995). More recently, Loughin and Noble (1997) have pursued the comparisons among available techniques while suggesting yet another method. Obviously, this is still a very open and active research field. The paper by Hamada and Balakrishnan is still another valuable attempt at establishing a comparison among these techniques by adopting their own approach, which we will discuss herewith.

The complexity of this problem stems for the multiple types of comparisons and hypotheses that can be made. One common such hypothesis is that of *factor sparsity.* It states that just a few of the considered factors have an influence on the response. Another widely accepted assumption is the normality of the noise distribution. In real-life, these assumptions may or may not be realistic. But, beyond that, the merits of a statistical method used in the identification of active effects is an elusive concept and has a much finer structure than the paper by Hamada and Balakrishnan may acknowledge. In the following discussion we will try to illustrate this point.

Assume that out of n measured effects there are j factors, which are truly active. That is, j factors are active and $n-j$ factors are not and can therefore be considered to be sampled values from the noise distribution. Of course, an ideal method would always find this result with probability 1. This would be "the truth, the whole truth and nothing but the truth". Now, what are the non-ideal alternatives? Rejecting the Null Hypothesis (no active effects exist) but for the wrong reasons is one case in point. In fact, there can be several such wrong reasons. The following is an exhaustive list of all the errors that can be made in this context:

Declaring

1. All j real factors as significant and, in addition, $1, 2, \ldots, n-j$ spurious effects as well,
2. Only some of the j real factors as significant and no spurious effects,
3. Only some of the j real factors as significant, plus $1, 2, \ldots, n-j$ spurious effects as well,
4. None of the j real factors as significant but identifying instead as significant $1, 2, \ldots, n-j$ other spurious effects,
5. None of the n effects as significant.

These five alternatives are either partially wrong or totally wrong. Notice that, from the Test of Hypothesis standpoint, all of these are alternatives to the Null Hypothesis and should be considered separately. The performance of the methods considered must be measured by taking into account how often they fall into one or the other of the above five wrong decisions.

Consider, for instance, declaring that a given drug treatment is safe and effective when in fact it is not or declaring that some treatments for weight loss are ineffective when in fact some of them may have a positive effect. These are not equivalent statements. What decision is preferable will depend on the circumstances.

Hamada and Balakrishnan have chosen to compare the different methods by measuring only their statistical power while fixing the number of experimental runs at 16 and tuning all the methods so that their IER is controlled at 0.044. (This is the average proportion of "false positives" for the particular method being considered when no real effects exist.) However, the proportion of false positives when there are some real effects is also important. In addition, this proportion may vary depending on the actual number and size of these effects. Clearly, this problem is compounded by the fact that it is impossible to consider each and all the possible situations that can arise in all the factorial experiments. Benski and Cabau (1995) suggested a *Figure of Merit* Q, which summarized some desirable properties of these techniques. Although this Q factor was far from perfect, we felt it conveyed more of the story than just statistical power and IER (or EER). Hamada and Balakrishnan in their Discussion recognize some of these points and give suggestions for further work in this area. One may add robustness to deviations from the assumptions as another characteristic worth investigating. In addition, the extension of these measurements to other common sample sizes, 8 and 32 come to mind, would have been valuable.

We fail to understand why Hamada and Balakrishnan claim in their Discussion that other publications testing methods for unreplicated experimental designs "have only compared their performance ... using some data sets", implying that no simulation tests were performed by others. Benski and Cabau (1995) clearly state that samples were Monte-Carlo generated for their tests.

Another important point in which we differ from the approach of Hamada and Balakrishnan is in the treatment of active effects. For their simulation experiments, they considered active effects as *fixed*. Although this is not wrong, we think that a more realistic approach is to consider active effects as *random* values issued from a scale enlarged distribution with respect to the noise distribution, as suggested by Box and Meyer (1986). In the *random* effects approach only the scale-shift is fixed.

In spite of these shortcomings, we view the paper by Hamada and Balakrishnan as a worthy addition to the ongoing efforts at better understanding the decision-making process in unreplicated experimental designs.

Acknowledgement

This author would like to thank Dr. Yvon Rebière for many helpful comments.

Schneider Electric, DRD-A2, 38050 Grenoble Cedex, France.
E-mail: claudio_benski@mail.schneider.fr

COMMENT

Perry D. Haaland

Becton Dickinson Research Center

The authors have made an important contribution to the literature by providing a comprehensive and equitable comparison of the wide variety of tests for active effects in effect saturated fractional factorial designs. The explanations of the methods are clear, excellent recommendations are made for new methods and extensions of existing methods, the comparisons are well justified, and the conclusions are clear. A key component of the informative nature of this study is that the authors have adjusted for common individual error rates (IER) in a sensible way. They have also proposed and used a common notation for describing the various methods and their properties. This sets a high standard for future work, and referees and editors should hold future authors to this standard.

Given this excellent start, the time now seems right for those of us in industry, who stand the most to benefit from the continued evolution of this technology, to find a means of supporting and extending the approach taken by the authors. I would like to propose an industry financed project for this purpose. This project could support an archive of software for the standardized comparison of

new methods. Clear guidelines could be given for this evaluation. Hopefully the authors could begin this effort by contributing their own code as the start of the software archive. Then new code could be submitted to evaluate any new method in comparison to the existing library of procedures. It would be highly advantageous if the archive could be organized so that simulations of existing methods would not need to be repeated, and so that it would be easy to evaluate all methods already in the library according to new criteria as they arise. It would be highly desirable if this project could also support work into the development of better, more interactive graphical displays for the analysis of effect saturated designs. I would like to invite anyone interested to contact me directly (pdh@bdrc.bd.com) if they would like to support or participate in this effort.

Since there is little to criticize or improve about this paper, let me instead espouse a few opinions regarding the general problem of testing for active effects in saturated fractional factorial designs. First, in my opinion, a purely statistical (that is, objective) identification of the active effects is not possible. The "vital few" and "trivial many" will in general be obvious regardless of the test method used. The "statistical in-between" effects will always be a problem (as the authors note) due to a combination of lack of power and the absence of an omnibus test that works over a wide range of the number of active effects. In addition, this is not purely a statistical problem as the physical interpretation of the marginal effects will always be critical to the analysis; for example, see Carlson (1992), p155 ff in which the largest marginal effect has no sensible interpretation.

So why does anyone use effect saturated fractional factorial designs? In my opinion, all experimenters are Bayesian, otherwise they would not be willing to use these small designs. This implies a substantial combination of subject matter knowledge and previous experience that they bring to bear on the problem. Does this mean that the experimenter has a well defined prior on the number of active effects that can be used to formulate the best test? I think not, because the experimenter's priors are generally quite complex – different factors have different likelihoods of being active or interacting with other factors. In addition, priors also get transformed to posterior distributions in a complex way. Consequently, it does not seem to be sufficient to use a method that relies too heavily on the assumption of factor sparsity with a fixed alpha. The choice of alpha = 0.20 by Box and Meyer (1986) is probably subject to a publication bias. My experience is that alpha is usually greater than 0.2 and often as high as 0.4. I think that the practical use of Bayesian methods is to vary the prior on alpha until the posteriors seem "reasonable" or to estimate alpha and k from the data (empirical Bayes). The Bayesian approach to analysis, however, has not seemed to be that useful in practice.

Whether using a frequentist or Bayesian approach, you cannot avoid the problem that the best test depends on knowing the unknown results; that is, the number of active effects. When there are only a few important effects, then any of the direct methods are going to be more or less equivalent with methods based on efficient estimation of sigma having a slight edge. When there are an intermediate number of active effects, the authors identify several tests that are appropriate. Based on my experience in practice, I recommend the use of the pseudo standard error method LEN89 as modified by Haaland and O'Connell (1995) to add some power for a moderate number of effects without much cost when effects are sparse. There is no clear choice for large numbers of effects. Given this state of affairs, the experimenter derives the most value from choosing a good design and then from interacting with graphical displays of the results. Consequently, there is no practical means to evaluate test performance in use because it becomes a subjective function of the experimenter's interpretation of the results and of the true number of active effects.

There is some advantage, then, to thinking about what the authors' results imply for the graphical analysis of effect saturated designs. A key issue is how the experimenter (often a nonstatistician) is going to interact with the analysis/software when interpreting the results. Thus, great value is added if the test results can be readily displayed; for example, by drawing a line on the Pareto plot of the absolute values of the effects (Haaland and O'Connell (1995)). This can be done with any of the direct tests. In general, I prefer the Pareto plot to half normal plots because there are fewer complex visual comparisons to make. (I concur with the authors in their recommendation against the use of full normal plots for this purpose.) For example, is the line sensible? Which effects depart from the line? You do, however, lose the information in the y-axis when doing the Pareto plot. The method for redefinition of the reference line presented by Johnson and Tukey (1987) is quite interesting in this regard. It is not clear, however, that the inherent variability of the half normal plot is improved by this method. While methods such as Zahn (1975a) may derive from the half normal plot, the results are still probably best displayed on the Pareto plot.

Given the lack of a single best test, it seems worthwhile to allow for the use of multiple lines in the Pareto plot wherein each line corresponds to a different test method and the test methods are chosen to provide good performance over a range of numbers of active effects. Given the lack of a clear choice for a best test and the importance of graphical display, we would do ourselves a disservice by relying too much on formal testing and P-values. Consider the role that two factor interaction plots play in the identification of important effects and in the choosing of directions for follow up experiments (Haaland (1989)). If there are any two factor interactions among the marginal or obviously significant effects,

then most experimenters need to look at interaction plots in order to evaluate what to do next. For one reason, the estimated effects correspond to a particular set of contrasts and careful study of an interaction plot may reveal that there are other more interesting or important contrasts to consider. There is in general no need to go back and estimate these contrasts as the experimenter can evaluate them graphically. Interestingly, this procedure is tied in with the assessment of significant effects because interaction plots are most useful when there are confidence bars shown. These confidence bars are generally derived from an estimate of sigma based on the smallest effects. Testing of effects is most meaningful in connection with an examination of the corresponding interaction plots, and this adds further complication to the meaningful evaluation of competing test methods.

Another complication in the comparison of these methods arises from how these designs are used in practice. In particular, the objectives of the experiment are generally to move toward the best process settings while at the same time deciding which factors are important. This is a composite objective, but it is only easy to evaluate the methods based on the second part of the objective. In fact, the experimenter often needs to compromise on the second part of the objective in order to better satisfy the first part. If null effects are misidentified as being active, the cost to the experimenter will primarily be in including excess factors in a follow-up experiment, as long as all of the important factors have been identified. The cost of failing to find an important factor is almost certain failure to optimize the process. Most experimenters would be willing to having a higher Type I error rate in order to more quickly and surely optimize the process under study, which leads us into the discussion of error rates.

My experience suggests that the power to detect individual important effects (and by implication the ability to optimize the process) suffers severely when the experimentwise error rate (EER) is tightly controlled. Consequently, an experimenter is seldom going to care very much about the EER, and to the extent that control is important it is better to focus on the individual error rate (IER). I have been uniformly disappointed with the performance of sequential tests and iterative methods, and I believe this is because sequential methods control the EER rather than the IER. Furthermore, IER will generally be quite small for direct methods when there is at least one large effect (we have empirical results for tests based on the PSE). As it is rare not to have at least one important effect, a good recommendation for process optimization would be to use methods that control the IER but choose a value of alpha greater than 0.05.

Equitable comparison of the methods is critical and the authors have chosen quite sensibly to control the IER in comparing tests. Other approaches could also be taken. For example, the power could be fixed for all methods and then

the IERs could be compared for the null model and for the second largest effect. I think that this would also lead to the practical recommendation that IER values greater than 0.05 should be used.

Whether or not an appropriate transformation is used can be more important than the test selected. The same is true for the impact of outliers. It is good to see that slight departures from normality do not affect the distributions of the test statistics too much, but of course, outliers do impact the sizes of estimates of the effects so their effective detection and elimination is critical to successful process optimization. More work needs to be done on this problem. In particular, it would be wonderful if interactive graphical methods could be developed for this problem.

In conclusion, fractional factorial designs seem destined to remain an important tool in an experimenter's toolbag for sometime to come. Given restrictions on time and cost, unreplicated designs will also continue to be widely used. There are many competing test methods, which are difficult to distinguish among based on performance, and it is easy to come to wrong conclusions if a careful comparison is not made. The choice of a test is further complicated by the fact that there is no one test that performs well over a wide range of numbers of active effects. In this regard, we need to ask software designers to provide us with an appropriate choice of tests with well controlled IER that are implemented within the context of a graphical analysis environment. I highly recommend that future authors take the comprehensive approach of this paper as a guideline for the evaluation of new methods, and that those who benefit the most from the development of new methods should help finance more practical evaluations and refinements, especially graphical displays.

Becton Dickinson Research Center, P.O. Box 12016 / 21 Davis Drive, Research Triangle Park, NC 27709, U.S.A.

E-mail: pdh@bdrc.bd.com

COMMENT

Russell V. Lenth

The University of Iowa

Hamada and Balakrishnan are to be congratulated for bringing together the sizeable literature on the subject of unreplicated experiments, and for their efforts to classify the various methods and to compare them fairly.

In this discussion, I have a few comments on the scope of the study and on its practical implications. Then I offer some ideas regarding the capability of an experiment (and of a procedure), and the kind of investigation I would like to see in the future.

1. Inference Space

In evaluating Monte Carlo results, it is important for the reader to understand that a Monte Carlo study is an experiment; and, like all experiments, one should take a careful look at the inference space. To what situations, exactly, do the results apply? And how generalizable are the results to situations not covered in the study? Due to practical constraints, a Monte Carlo study is often very limited in scope, and this one is no exception.

This particular study considers only cases where there are 15 independent effects (based on $n = 16$ observations). I feel that that is *not* a serious limitation. It is appropriate to keep the focus on behavior for small amounts of data because that is where these methods apply. The authors comment (based on another study of theirs) that it is hard to make any inferences with only 7 effects. One can guess that the methods perform comparatively the same as the number of effects increases above 15. That is, 15 effects are probably enough so that the "true" properties of the methods themselves have "kicked in". Thus, there is a comfort level in believing that the study makes a reasonable comparison among the methods.

Another limitation, perhaps more serious, is that in the power comparisons, all active effects are of the same magnitude. Again, however, it seems reasonable that power would be some continuous function of effect size and effect mix. It is possible, however, that different methods could compare differently depending on the mix of active effects of different sizes. This seems particularly likely in comparisons of different general classes of methods (i.e., composite, directed, sequential).

The limitations in the robustness area are more severe and are worth noting carefully. The authors do consider the case where the data have standardized $t(9)$ errors; however, this is not a particularly heavy-tailed distribution, especially when 16 such observations are averaged together (as the authors show in their Figure 6). Moreover, it could very well be true that a skewed distribution of errors could cause as much or more havoc than a heavy-tailed one. So I would say that we still do not know very much about robustness.

2. Pick Your Favorite

The most striking result to me is the degree of sameness of power of most of the methods being compared—particularly the directed ones. If a directed procedure is desired, there is little to lose in picking the one that is most convenient, aesthetically pleasing, or whatever.

There is a limited amount of information available in an unreplicated experiment, and apparently these analyses make good use of it. For that reason, I forecast that there will be no breakthrough in this area of research, where some new method of analysis is found that greatly outperforms the ones already studied here. (Nevertheless, I thank the authors for suggesting guidelines for testing future methods.)

I am also happy to see the directed methods as being so successful in competition with the sequential and composite methods. I like the directed methods because they go straight to the point of comparing effects with a standard. They are easy to explain, and can be displayed using a Pareto chart or similar graph. It is important to remember that our analyses have an audience, and (assuming that it is a good analysis) the simpler it is to explain and to present, the better.

3. Capability of an Experiment

For convenience in this discussion, let me define the "decent-chance detection capability"—DCDC(α)—to be the effect size that can be detected with 50% power when the size of the test is α. In Figure 4 we see that if $n = 16$ and there are 4 active effects, then DCDC(.044) is at least 3σ for most procedures. When there is only one active effect, the DCDC reduces to about 2.5σ.

I like having this quantitative result available. Coincidentally, a DCDC of 3σ is a particularly useful point of conversation when discussing the capability of a proposed experiment with engineers and managers. That's because the most popular measures of process capability are also based on 3σ.

In a control chart, the control limits are also at "$\pm 3\sigma$", but this refers to the standard deviation $\sigma_{\overline{x}}$ of the batch mean. For batch sizes of size 5 (very popular), $3\sigma \approx 6.7\sigma_{\overline{x}}$—far beyond the control limits. In short, 3σ is no small effect, and in particular an effect of much smaller magnitude could throw a process seriously out of control.

Now, let's look at this from a practical angle. These small unreplicated experiments are usually used for screening; they are seldom meant to serve as definitive scientific experiments. Instead, a screening experiment is only one part of a process of experimentation where the results of one experiment guide the design of the next. In a screening experiment, the consequence of making a type I error would be doing the wrong experiment in the sequel—maybe not that serious a mistake. Perhaps it is unrealistic in this context to even think about testing effects at $\alpha = .044$ or .05. An α of .20 may be quite suitable; and DCDC(.2) would be closer to the 1σ–1.5σ range of effect sizes that one would want to detect. (This said, I necessarily have to be comfortable with a *really* large EER in a screening experiment. That says that we are guaranteed to chase down some blind alleys once in awhile in the sequence of experiments.)

It seems useful to develop some good ways of describing and quantifying the capability of an experiment. We could in fact use DCDC(.2) for that purpose. But in light of the above discussion, I suggest as an alternative the following: Let $C_{k,m}$ denote the value of α such that DCDC(α) = $m\sigma$ when there are k active effects of size $m\sigma$. It is a measure of risk, so that a small $C_{k,m}$ is desired. For practical purposes, $C_{1,1.0}$ seems most appropriate, but some may prefer to raise the number of active effects (e.g., $C_{4,1.0}$). The same capability measure could be used to compare different procedures, an alternative to the one cited in Benski (1995) (I have not seen that paper).

4. An Idea for the Future

I have already said that I am not holding my breath for any new procedures that will outperform the ones already studied. What I would like to see instead would be studies that in some way simulate the process of iterative experimentation where the results of one experiment determine the design of the next. For example, suppose that there are 30 factors available to experiment on, but that we can do screening experiments of $n = 16$ runs (so that at most 15 factors can be investigated in one experiment). Given that there are, say, 4 active main effects and 6 active two-way interactions, how many experiments will it take to identify at least 8 (or all 10) of these effects?

It sounds like a real challenge to develop a realistic model for such a process of iterative experimentation but any success in doing so would shed new light on how these procedures compare in practice. We would also want to compare the procedures with unsophisticated strategies, such as "always incorporate the l largest absolute effects in the next experiment". The procedures might compare very differently in such a setting. I'm guessing that there would be a few surprises.

The University of Iowa, Department of Statistics and Actuarial Science, Iowa City, IA 52242-1419, U.S.A.

E-mail: rlenth@stat.uiowa.edu

REJOINDER

M. Hamada and N. Balakrishnan

We thank C. F. Jeff Wu, the previous Chair Editor, for organizing this discussion. We also thank the discussants for providing thought provoking comments on a variety of issues.

Comparing the Methods

The discussants suggest additional criteria for evaluation. Benski points out that a single criterion such as power need not tell the whole story; he suggests looking at IER (i.e., the average number of inert effects declared active) *when some of the effects are active*. We captured these results in the study although we did not report them; they were consistently below 0.044 which suggests there is an opportunity for devising methods that are less conservative and therefore more powerful. Lenth proposes using a criterion based on the "decent-chance detection capability" (DCDC). We like the practical importance of DCDC for conveying a proposed experiment's capability to experimenters and their managers; experimenters will know the magnitude of the effects that have a decent chance of being detected by the experiment.

The discussants make several remarks about the scope of our simulation study. Lenth suggests that there might be more differentiation between the methods under scenarios with unequal magnitude active effects; this needs more study. Benski suggests using a random effects approach for generating the active effects – from the model specified in Box and Meyer (1986), the number of active effects varies from experiment to experiment and the magnitudes of the active effects are also random. Thus, evaluation under this approach can be interpreted as the performance under all possible experiments where the average number of active effects and average magnitude of the active effects are specified. Here, perhaps there will be less differentiation between the methods because evaluation is done over a mixture of "fixed effects" situations; i.e., a method will do better for certain fixed effect configurations but not for others. Regarding distributional assumptions, Lenth raises the possibility of skewed errors which is related to Haaland's issue of transforming the data. For example, reliability data tend to be skewed and are modeled by the exponential or Weibull regression models; for these models, the log data follow a location-scale model as is assumed by the methods for unreplicated experiments except that errors are no longer normal but skewed. Thus, the data first need to be transformed before estimating the factorial effects! The skewness of the errors should be ameliorated by the central limit theorem effect since the estimates are linear combinations of observations, but this needs to be explored further.

Using the Methods in Practice

Haaland and Lenth note that a primary context for small unreplicated experiments is screening. Here, because it is more serious to miss important factors (which would not be studied in subsequent experiments) than to misidentify unimportant factors, both have no qualms with using an IER exceeding 0.05. (Consequently, neither are concerned with the high EER that results.) Lenth

chooses an IER of 0.20 because the methods have a decent chance of detecting active effects in the 1σ-1.5σ range. It is worth commenting that the experimental "σ" (which depends on the way the experiment was carried out) may be much smaller than the process σ; thus, the methods using an IER as small as 0.05 would still have a decent chance of detecting 3"σ" sized effects, where 3"σ" could be as small as 1σ. This suggests that the low power results for small sized effects from our simulation study need not be viewed so disappointingly.

Haaland emphasizes the importance of graphical methods for helping the experimenter to interpret the experimental results. He suggests using a vertical plot of the absolute values of the estimated effects with a horizontal line (associated with one of the objective methods) drawn where the important effects are easily identified as those exceeding the line. Several lines might be drawn corresponding to a different methods and/or different IER values. We heartily agree. The guidance provided by the objective methods is needed, especially for 8 run experiments.

Future Research

To clarify a misunderstanding raised by Benski, we had noticed that *some* papers had proposed new methods and had only demonstrated their use with some data sets. Our point was that such demonstrations alone do not provide a sufficient evaluation of a new method. Accordingly, we made suggestions for evaluating new methods. We are happy to see that Loughin and Noble (1997), who had a copy of our technical report, followed our recommendations. Notably, Loughin and Noble (1997) propose a nonparametric method that performs competitively with Lenth's (1989) method for a small to moderate number of active effects and that performs amazingly well for up to 10 active effects (out of 15)!

Lenth issues a tantalizing challenge: evaluate a procedure or more likely a suite of procedures in an iterative experimentation process, where the results of one experiment determine the design for the next. Haaland's remarks about the impact of two-factor interaction plots on subsequent experimentation is also relevant. For such an evaluation, an integral component is the experimental designs employed. For example, screening designs need not be restricted to *regular* two-level fractional factorial designs. Lenth notes that the regular 16 run design can screen up to 15 factors, but there are 12 run supersaturated designs that can handle up to 66 factors. The estimated effects are no longer independent so that the methods discussed here are no longer applicable. (See Chipman, Hamada and Wu (1997) for an appropriate analysis methodology.) Experimental designs also need to be suitably chosen in subsequent stages. We look forward to seeing this research pursued.

In conclusion, we fully agree with the suggestion made by Haaland that an industry-funded repository of methods for analyzing unreplicated experiments be created. We are certainly interested in participating in such a venture. We would also like to encourage industry to support and participate in research pertaining to experimental design beyond unreplicated experiments such as that raised by Lenth's challenge. We believe that such a collaboration will only result in benefits for both parties concerned.

Additional References

Carlson, R. (1992). *Design and Optimization in Organic Synthesis.* Elsevier Inc., New York.

Chipman, H., Hamada, M. and Wu, C. F. J. (1997). A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics* **39**, 372-381.

Haaland, P. D. (1989). *Experimental Design in Biotechnology.* Marcel Dekker Inc., New York.