Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals

N. Balakrishnan and M.S. Hamada
University of Waterloo

RR-94-12

November 1994

ANALYZING UNREPLICATED FACTORIAL EXPERIMENTS: A REVIEW WITH SOME NEW PROPOSALS

N. Balakrishnan
Department of Mathematics and Statistics
McMaster University

M. Hamada
Institute for Improvement in Quality and Productivity
University of Waterloo

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 focus of this paper is a comparison of these methods and their variants via an extensive simulation study. Many methods are comparable, but clearly some cannot be recommended. The results from the study suggest some recommendations 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: 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 so called "active" effects which are non-zero.

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) and Schneider, Kasperski and Weissfeld (1993). 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" (i.e., zero) 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 various 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; e.g., 7, 15, 31 for the commonly used designs with run sizes n of 8, 16 and 32. 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 error variance is σ^2, so that the 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. (Even if the errors are non-normal, the contrasts are nearly normal by the Central Limit Theorem.) Finally, the size of the active κ_i will be given in multiples of σ, the 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 (1993). The problem in comparing these methods is to do so on an equal basis since they perform differently when all the effects are inert. Thus, a cali-

bration is needed which we do by considering sequential procedures based on the particular statistics used by the various methods. That is, the largest unsigned contrast is tested to determine whether the corresponding effect is active. Then the next largest unsigned contrast is considered and so forth. Thus, the goal of the comparison is to identify which statistics inherently provide better performance. Performance of the procedures is evaluated through an extensive simulation study.

In Section 5, the paper presents some requirements 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.

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 above the line, correspond to active effects. Note the implicit assumption of few active effects in order to draw a line through the 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 deviation 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

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 deviation
κ_i	ith effect
c_i	ith contrast - estimate of ith effect
$\lvert c\rvert_{(i)}$	ith order statistic of the unsigned contrasts
M_i	mean square based on c_i
$M_{(i)}$	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 of all contrasts c_i
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

List of Methods

DAN59	Daniel (1959) uses (1)
HB69	Holms and Berretoni (1969) uses (2)
ZAHN75	Zahn (1975a, version S) uses (4)
ZAHN75(m)	uses m smallest contrasts to estimate τ
STUK82	Seheult and Tukey (1982) uses (5)
BM86	Box and Meyer (1986)
JTUK87	Johnson and Tukey (1987) uses (7)
LEN89	Lenth (1989) uses (11)
BEN89	Benski (1989)
BIS89	Bissell (1989)
BP91	Berk and Picard (1991) uses (14)
HLOH92	half-normal version of Loh (1992)
JP92	Juan and Pena (1992) uses (21)
DONG93	Dong (1993) uses (22)
SKW93	Schneider et al. (1993) uses (23)
MSKW	SKW93 accounting for k'
MLEN	LEN89 accounting for k'
MDONG	uses iterative DONG93 estimator
CORR	correlation coefficient probability plot uses (24)
HSW	half-normal Shapiro-Wilk uses (26)
DISP	dispersion test uses (27)

absolute values. Note how the unknown scale is removed by the standardization and that only the largest four of the 15 unsigned contrasts (about 25%) can be tested sequentially starting with the largest. The guardrails drawn on the plot are the corresponding critical values for the modulus ratios (1). 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. Zahn (1975a) pointed out problems with the guardrails but this need not concern us.

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:

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

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 the smallest 68.3% of the 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 square error 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. 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 version.

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 $2^{k+1} - 1$ synthetic contrasts. 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.

Note the similarity with Benski's (1991) outlier test to be discussed later which uses d_F the interquartile range of the original contrasts c_i (which does not include the additional zero contrast). d_F which is a robust estimate of spread provides a basis for estimating τ (Juan and Pena 1992). Since d_F estimates $2\Phi^{-1}(0.75)\tau$, then $d_F/(2\Phi^{-1}(0.75)) = d_F/0.7413011$ estimates τ. Consequently, in the spirit of Daniel (1959) and Zahn (1975a), the estimate denoted by $\hat{\tau}_{\mathrm{DF}}$ could be used to standardize the unsigned contrasts:

$$|c|_{(i)}/\hat{\tau}_{\mathrm{DF}} \tag{5}$$

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 have a $N(0, \sigma^2_{active})$ distribution. Letting $K^2 = (\sigma^2_{inactive} + \sigma^2_{active})/\sigma^2_{inactive}$, then the contrasts c_i follow a $(1-\alpha_{active})N(0, \sigma^2_{inactive}) + \alpha_{active}N(0, K^2\sigma^2_{active})$ distribution. 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 were (0.13-0.27) and (2.7-18) with averages of 0.2 and 10, 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.,

$$|c|_{(i)}/|\tilde{z}|_{(i)k}, \tag{6}$$

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:

$$\text{ratio-to-scale} = \text{display ratio/median display ratio.} \tag{7}$$

Critical values given in their Table 12 are for the maximum ratio-to-scale in a sample of size k. 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:

$$f(|c|_{(i)})/\sum a_i f(|c|_{(i)}), \tag{8}$$

for some constants a_i. Note that (1), (2), (4) and (14) to be discussed later fall in 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.

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 consistent for τ (as k $\rightarrow \infty$) 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):

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

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

$$W' = (\sum_{i=1}^{k} z_{(i)k}c_{(i)})^2/(\sum_{i=1}^{k} z_{(i)k}^2 \sum_{i=1}^{k}(c_{(i)} - \bar{c})^2), \tag{12}$$

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 (12) can also be viewed as a correlation-type statistic (i.e., the mean of $z_{(i)k}$ is nearly zero), a large value (close to one) indicates a strong association between the expected normal order statistics and the ordered 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.

Because this procedure is a hybrid of two tests it cannot be directly compared with the other methods which use a single statistic. A comment about the first test 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. The outlier test based on d_F will be studied in terms of the form (5) discussed previously.

Bissell (1989, 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:

$$S_M/\bar{M}, \tag{13}$$

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 (1989) suggested dropping several mean squares that are obviously active and then retesting the remaining effects. Note that the χ^2 approximation does not account for the fact that the estimate $\bar{M}$ is used rather than the true mean.

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:

$$M_{(l)}/\sum_{i=1}^{m} M_{(i)}. \tag{14}$$

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.

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 there is a computational drawback to finding the set of contrasts used in the normal plot. Take for example, a Plackett-Burman 12-run design with 11 factors, which would require 2^{11} sets of contrasts to be considered. Working with the unsigned contrasts eliminates all this computation, however. This suggests using a half-normal version which will be considered in Section 3.

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. 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 mean square error 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.

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: $s_{\text{DONG}} = \sqrt{m^{-1}\sum_{\{|c_j|<2.5s_0\}} c_j^2}$, 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 mean square error than PSE which provided his motivation for using it to standardize the contrasts as

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

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, Kasperski and Weissfeld (1993)

Schneider, Kasperski and Weissfeld (1993) proposed standardizing the contrasts by an estimator of τ given in Wilk, Gnanadesikan and Freeny (1963). By treating a set of the 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) $\hat{\tau}_{\text{CEN}}$. 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:

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

3. Modifications and New Proposals

Some modifications and 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, $k' = \text{ceiling}[k/4]$, S_2 is the root mean square 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.

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 $\tilde{z}_{(i)k}$.

$$R_{\text{CORR}} = \sum_{i=1}^{k}(\tilde{z}_{(i)k} - \bar{\tilde{z}})(c_{(i)} - \bar{c}) / \left(\sqrt{\sum_{i=1}^{k}(\tilde{z}_{(i)k} - \bar{\tilde{z}})^2}\sqrt{\sum_{i=1}^{k}(c_{(i)} - \bar{c})^2}\right). \tag{24}$$

Note the similarity with the modified Shapiro-Wilk statistic W' in (12) 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 (24) above. Small values of R_{CORR} suggest the presence of active effects.

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)

$$\hat{\tau}_{\text{BLUE}} = \underset{\sim}{\mu}^T \Sigma^{-1} |\underset{\sim}{c}|_{(\,)} / (\underset{\sim}{\mu}^T \Sigma^{-1} \underset{\sim}{\mu}), \tag{25}$$

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 (25) for n=8 and n=16, respectively. Since the MLE of τ based

on the $|c_i|$ values is

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

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

$$\text{HSW} = \hat{\tau}_{\text{BLUE}}/\hat{\tau}_{\text{MLE}}, \tag{26}$$

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. Analogous to the Shapiro-Wilk test, the critical region was taken to be small values of HSW which was 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

$$D_m = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{|c|_{(i)}}{PSE\ |z|_{(i)k}} - 1\right)^2. \tag{27}$$

Note that since PSE is a "robust" estimator of τ, a significant departure of $|c|_{(i)}/\text{PSE}$ 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 (27). This statistic can also be used sequentially.

4. A Comparison 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) and Haaland and O'Connell (1993). 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 when all

effects are inert: DAN59 for Daniel (1959), HB69 for Holms and Berretoni (1969), ZAHN75 for Zahn (1975a, version S), STUK82 for Seheult and Tukey (1982), BM86 for Box and Meyer (1986), JTUK87 for Johnson and Tukey (1987), LEN89 for Lenth (1989), BEN89 for Benski (1989), BIS89 for Bissell (1989), BP91 for Berk and Picard (1991), HLOH92 for the half-normal version motivated by Loh (1992) and presented in Section 3, JP92 for Juan and Pena (1992), SKW93 for Schneider et al. (1993) and DONG93 for Dong (1993). Based on 10,000 simulations for a 16 run experiment (k=15), Table 1 gives the observed proportion of simulations that zero to eight effects were declared active when all effects were inert. Note that no two procedures have exactly the same performance.

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 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, $\sum(i/k)p_i$. This definition of IER when all effects are inactive can be extended to the case when some effects are active by suitably changing k 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 use the IER or EER criteria. DAN59 (critical values from Zahn 1975a) and ZAHN75 attempt to control IER at 0.05. Note that DAN59 can detect at most four effects. Also, ZAHN75 was used to detect at most four effects here. HB69 was started by pooling the nine smallest effects and used α_{pool}= 0.25 and α_{final}=0.05 so that IER at 0.05 is implied. LEN89, BIS89, 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

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 $\geq$ 8 declared effects)

	number of declared effects										
method	0	1	2	3	4	5	6	7	8	IER	EER
DAN59	.598	.193	.093	.050	.065					.0527	.402
HB69	.629	.157	.067	.045	.033	.029	.042			.0634	.371
ZAHN75	.618	.190	.089	.048	.055					.0487	.382
STUK82	.742	.129	.054	.026	.017	.012	.008	.005	.005*	.0387	.258
BM86	.748	.176	.044	.016	.007	.004	.003	.002	.000	.0262	.252
JTUK87	.950	.034	.010	.003	.002	.001	.001	.000	.000	.0054	.050
LEN89	.755	.144	.054	.024	.013	.007	.003	.001		.0290	.245
BEN89	.952	.037	.008	.002	.001	.001	.000			.0037	.048
BIS89	.834	.118	.032	.011	.004	.001	.000			.0157	.166
BP91	.555	.259	.119	.050	.017	.004	.000			.0492	.445
HLOH92	.951	.017	.018	.010	.004	.001	.000			.0070	.049
JP92	.799	.104	.039	.021	.014	.010	.006	.004	.003*	.0294	.201
SKW93	.590	.254	.105	.038	.011	.002	.000			.0421	.410
DONG93	.569	.302	.085	.029	.011	.004	.001	.000		.0418	.431

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 controls 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).

The challenge is then to compare these methods on an equal basis. Note that the essence of most of these methods except the hybrid ones (Benski 1989, Loh 1992) is a single statistic that can be used for deciding whether the effect corresponding to the largest unsigned contrast is active or not. This statistic can then be used sequentially to consider the next largest unsigned contrast and so forth. The resulting sequential procedures using the various statistics can then be calibrated so that all have the same performance when all effects are inert, i.e., the same p_i. Thus, the main objective of our comparative study is to identify which statistics inherently perform better when there are active effects. The study will focus on designs with n=8 and n=16 runs (i.e., k=7 and k=15).

DAN59 (1), ZAHN75 (4), JTUK87 (7), LEN89 (11), JP92 (21), SKW93 (23) and DONG93 (22) which compute standardized contrasts already have the form of a single statistic. STUK82 can be adapted to have the same form using (5) based on the interquartile range, d_F. Recall that this is related to the second test of Benski (1989) and the first test of Loh (1992). Note that DAN59, LEN89, JP92 and DONG93 use a fixed denominator. The denominator for ZAHN75(m) is based on the m smallest unsigned contrasts but the coefficients $|z|_{(i)k'}$ depend on k', the current number of contrasts being considered. For k=15, DAN59(9) which standardizes the contrasts by $|c|_{(9)}$ will be used since DAN59(11) can only detect at most four effects; consequently, the performance of DAN59(11) with all inert effects will not be comparable. ZAHN75(11) can detect more than four effects although its power is expected to be significantly diminished. Similarly, DAN59(4) is used for k=7. Modified

versions of SKW93 and LEN89 will also be studied which account for k', the current number of contrasts being considered, so that their denominators are variable. MDONG refers to the procedure which uses the iterative estimate of τ based on (22) proposed by Dong (1993). Note that the denominator for MDONG is fixed.

HB69 (2) and BP91 (14) compute standardized mean squares so that they also have this form of a single statistic. Recall that HB69 starts out by pooling the m=4(=9) smallest mean squares when n=8(=16) and possibly pools additional mean squares. In contrast, BP91 pools the m=4(=9) smallest mean squares when n=8(=16) in a fixed manner.

BM86 can be adapted using the marginal posterior probability of the largest unsigned contrast as the single statistic.

All the methods discussed thus far focus on the largest contrast directly which will be referred to as directed tests in the following. Contrast this with BIS89 which is designed to detect the presence of one or more active effects and which we refer to as a composite test. It is not specifically directed at the largest contrast but nevertheless can be used sequentially as proposed. The new composite test statistics DISP (27), CORR (24) and HSW (26) presented in Section 3 can also be used sequentially. For HSW, m=k is used. Finally, a modified method based on Le and Zamar (1992) and denoted by MLZAM uses (19) in place of (17). MLZAM is also a composite test.

BEN89 and HLOH92 will not be considered further because they are hybrid tests and cannot be put in this form.

Some details of the calibration for n=8(=16) will be discussed next. For each of the methods listed above, critical values were chosen to ensure that $\mathbf{p}$=(0.6,0.24,0.096,0.064) for n=8 and $\mathbf{p}$=(0.6,0.24,0.096,0.038,0.015,0.006,0.004) for n=16, where $\mathbf{p}$ denotes the vector of p_i, the probability of declaring i effects active when all effects are inert. The summary measures EER = 0.40 and IER=0.088 for n=8 (=0.044 for n=16). Note that each method as implemented can detect at most three and six active effects, respectively. The critical values for the ith largest unsigned contrast were obtained by simulation and were

based on 10,000 samples whose statistics for all larger unsigned contrasts had exceeded their corresponding critical values. The critical value is then found by identifying the 0.60 percentile of the 10,000 statistics. For n=8, the empirical 0.50 and 0.60 percentiles of JTUK87 test statistic for the largest unsigned contrast were the same because of its discreteness, so that it could not be calibrated. Consequently, JTUK87 was not included in the study for n=8. Also, for n=16, BM86 could not be calibrated exactly because the simulation time required to calculate the critical values as discussed above is prohibitive. Instead, those effects whose marginal posterior probability exceeded 0.357073 were declared active; 0.357073 is the 0.60 percentile of the maximum marginal posterior probability of 15 contrasts under all inert effects so that EER is controlled at 0.40 like the rest of the procedures. Using the same threshold for testing the next largest unsigned contrast and so forth gave a $\mathbf{p}$=(0.602,0.258,0.081,0.030,0.012,0.007,0.005,0.004,0.001) which is surprisingly close to $\mathbf{p}$=(0.6,0.24,0.096,0.038,0.015,0.006,0.004) for the calibrated procedures.

Since the various methods are calibrated, their power can be investigated under various scenarios. This was also done by simulation based on 10,000 samples. For n=8, one to three active effects all having the same magnitude were studied whose sizes were from 0.5σ to 3σ. Recall that σ denotes the error standard deviation not the contrast standard error. For n=16, one, two, four and six active effects all having the same magnitude from 0.5σ to 3σ were studied. Figures 1a-c for n=8 and Figures 2a-d for n=16 display the power or average proportion of active effects that were declared active.

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

- All the methods are comparable for small size effects, say 0.5σ, which exhibit little power and depend little on the #AE and the run size n.

- Except for BIS89 (n=16, #AE=6), the power increases as the size of the active effects increases. Note that the effects need to be rather large relative to the error standard

deviation σ. For example, the power is around 0.7(0.45) for a single 1.5σ effect for n=16(=8).

- The power decreases as the number of active effects increases.

- Even in terms of equal proportions $\#AE/n$, the power is larger for n=16.

- The directed methods which focus on the current largest unsigned contrast tend to perform better than the composite methods, BIS89, MLZAM, DISP, CORR 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. MLZAM is an exception which performs surprisingly well, however. CORR is clearly the worst of all the composite tests.

- BIS89, a composite procedure appears promising say for 25% #AE but then its performance seriously degrades at 40% #AE. 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 (13). This is clearly an undesirable property.

- Most of the directed methods are similar. ZAHN75(4 and 9 for n=8 and 16, respectively), HB69, and BP91 have the best performance for #AE=3 and n=8 (#AE=6 and n=16) but recall that these methods assume at most three and six active effects when n=8 and 16, respectively. Note ZAHN75(11) for n=16 performs poorly when #AE = 6 so that ZAHN75(9) is preferable in this situation; ZAHN75(9)'s power for small #AE is reduced, but its loss of power is negligible.

- Most of the directed methods only differ in the estimator for τ. Various proposals were motivated by better mean square error 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 the estimator based on the d_F used

by STUK82 performed much worse than both $IMAD_0$ and PSE. SKW93 was motivated similarly with censoring being used to reduce the bias of the estimator. Finally, DONG93 used trimmed means instead of the trimmed medians used by PSE because of improved efficiency. Yet, the gains in estimator performance appear to have little impact on the test performance. For example, JP92 ($IMAD_0$) does not outperform LEN89 (PSE). DONG93, in fact does worse for large #AE; the threshold for active contrasts $2.5s_0$ has a greater impact on the mean square error of DONG93 (inflating it) than on the mean square error of PSE.

- MLEN and MSKW do not outperform LEN89 and SKW93, respectively. It appears that calculation using the current number of contrasts offers little if any improvement.

- 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.

- BM89 is competitive with the non-Bayesian directed procedures. It does somewhat worse for larger #AE, but recall that α_{prior} was set at 0.2, i.e., it was designed for 20% #AE.

Besides power, the IER or average proportion of inactive effects that were declared active as displayed in Figures 3a-c for n=8 and Figures 4a-d for n=16 needs to be studied. (Under all inert effects, the IER is .088 and .044 for n=8 and 16, respectively.) Note that only the three (six) largest effects are tested for n=8 (=16) so that caution is needed when comparing the results for different number of inactive effects. For example, Figures 3a and 3b show that IER does not decrease significantly as the size of the active effects increases; in fact, the IER of some of the procedures do not even behave monotonically. Contrast this with the behavior of the IER in Figure 3c which is clearly monotonic as the size of the active effects increases; here there are three active effects but only the three largest contrasts are being tested. Similar patterns may be seen in Figures 4a-d which display the IER results for

n=16. Note that among the methods with non-monotonic IER behavior, the performance of DONG and MDONG are clearly the worst. It is interesting that these methods are similar in principle to Lenth's (1989) method, which also suffers from the same non-monotonic behavior but to a much lesser degree.

Some other comments are:

- The results for HSW suggest that BEN89 will not be superior since it uses the Shapiro-Wilk test. Also, BEN89 works with the contrasts rather than the unsigned contrasts so that the arbitrariness of the factor level labels is not accounted for.

- BM86 requires a considerable amount of computation, but its performance is only comparable to the ZAHN75, HB69 and BP91, of which ZAHN75 and BP91 are especially simple to implement.

5. Discussion

Recently, several papers have proposed methods but 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 to evaluate its performance is needed to compare the new method with existing ones. The calibration problem needs to be dealt with 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. In particular, a nearly monotonic decreasing IER as the size of the active effects increases is desirable.

- 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, an EER of 0.40 was used in the simulation study which some might consider large, but yet the IER was 0.044 for n=16. Note that in the industrial setting there is a particular interest in not missing active effects which represent lost opportunities for improving quality. Also, this level of EER is conservative in light of the "fire-fighting" methods commonly employed in industry; i.e., a process change is made and if the quality characteristic is better than before the change (using no measure of variability), it is concluded that the process change has successfully improved quality.

- 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?

- Should other non-null scenarios be considered such as different sized active effects? If so, what would be appropriate choices for the sizes of the active effects?

- Are there non-sequential procedures which have better performance or are sequential directed tests preferable?

- Can gains be made by combining methods, i.e., the hybrid methods? For example, Loh (1992) indicated that the use of the prediction intervals for identifying active effects outperformed the use of outlier test based on the interquartile range which was used by Benski (1989).

- Can information on how many active effects there are likely to be present in the experiment be exploited?

- 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 interesting to explore how these varied works may help to suggest 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. Bear in mind that the study investigated a sequential version of most existing tests as well as the new ones proposed in Section 3. The recommendations are:

- Most of the directed methods perform similarly. Of these, the statistics for BP91 and LEN89 are particularly simple to calculate.

- The power for BIS89 seriously degrades when there are many active effects so that this method is not recommended.

- For the methods DONG and MDONG, the percentage of inactive effects identified as active depends non-monotonically on the size of the active effects. Because this is clearly an undesirable property, these two methods are 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 active effects and not misidentify inert effects when some effects are active. Consequently, objective methods, which

directly account for this variability, are preferable. Also for n=8, the real effects need to be rather large relative to the error standard deviation in order to be detected.

Acknowledgements

We thank Fred Hulting for kindly sending his FORTRAN program PSTPRB as described in Stephenson, Hulting and Moore (1989), and Jock MacKay and Clif Young for insightful comments. 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.

References

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

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

Benski, H.C. (1989), "Use of a Normality Test to Identify Significant Effects in Factorial Designs," *Journal of Quality Technology*, 21, 174–178.

Berk, K.N., and Picard, R.R. (1991), "Significance Tests for Saturated Orthogonal Arrays," *Journal of Quality Technology*, 23, 79–89.

Birnbaum, A. (1959), "On the Analysis of Factorial Experiments Without Replication," *Technometrics*, 1, 343–357.

Bissell, A.F. (1989), "Interpreting Mean Squares in Saturated Fractional Designs," *Journal of Applied Statistics*, 16, 7–18.

Bissell, A.F. (1992), "Mean Squares in Saturated Fractional Designs Revisited," *Journal of Applied Statistics*, 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*, 40, 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*, New York, NY: John Wiley & Sons, Inc.

Daniel, C. (1983), "Half-normal Plots," in *Encyclopedia of Statistical Sciences*, Volume 3, (Eds., S. Kotz and N.L. Johnson), New York, NY: John Wiley & Sons, Inc.

Dong, F. (1993), "On the Identification of Active Contrasts in Unreplicated Fractional Factorials," *Statistica Sinica*, 3, 209–218.

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," *Reports in Statistical Applied Research*, JUSE, 12, 149–164.

Gupta, S.S., and Panchapakesan, S. (1979), *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*, New York, NY: John Wiley & Sons, Inc.

Haaland, P.D., and O'Connell, M.A. (1993), "Inference for Effect Saturated Fractional Factorials," unpublished manuscript.

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

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

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 & Analysis*, Ed. C.L. Mallows, New York, NY: John Wiley & Sons, Inc.

Juan, J., and Pena, D. (1992), "A Simple Method to Identify Significant Effects in Unreplicated Two-level Factorial Designs," *Communications in Statistics-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," *Journal of Statistical Computation and Simulation*, 41, 41–54.

Lenth, R.V. (1989), "Quick and Easy Analysis of Unreplicated Factorials," *Technometrics*, 31, 469–473.

Loh, W.Y. (1992), "Identification of Active Contrasts in Unreplicated Factorial Experiments," *Computational Statistics & Data Analysis*, 14, 135–148.

Schneider, H., Kasperski, W.J., and Weissfeld, L. (1993), "Finding Significant Effects for Unreplicated Fractional Factorials Using the n Smallest Contrasts," *Journal of Quality Technology*, 25, 18–27.

Seheult, A., and Tukey, J.W. (1982), "Some Resistant Procedures for Analyzing 2^n Factorial Experiments," *Utilitas Mathematica*, 21B, 57–98.

Shapiro, S.S., and Francia, R.S. (1972), "Approximate Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, 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," *Journal of 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.

Voss, D.T. (1988), "Generalized Modulus-Ratio Tests for Analysis of Factorial Designs with Zero Degrees of Freedom for Error," *Communications in Statistics-Theory and Methods*, 17, 3345–3359.

Wilk, M.B., Gnanadesikan, R., and Freeny, A.E. (1963), "Estimation of Error Variance from Smallest Ordered Contrasts," *Journal of the American Statistical Association*, 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.

Appendix

Tables used to implement HSW (26) are given in this appendix.

Half-Normal Shapiro-Wilk Test

The statistic $\hat{\tau}_{\text{BLUE}}$ (25) 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

Figure 1a: Power for n=8
one active effect = .5(.5)3 sigma

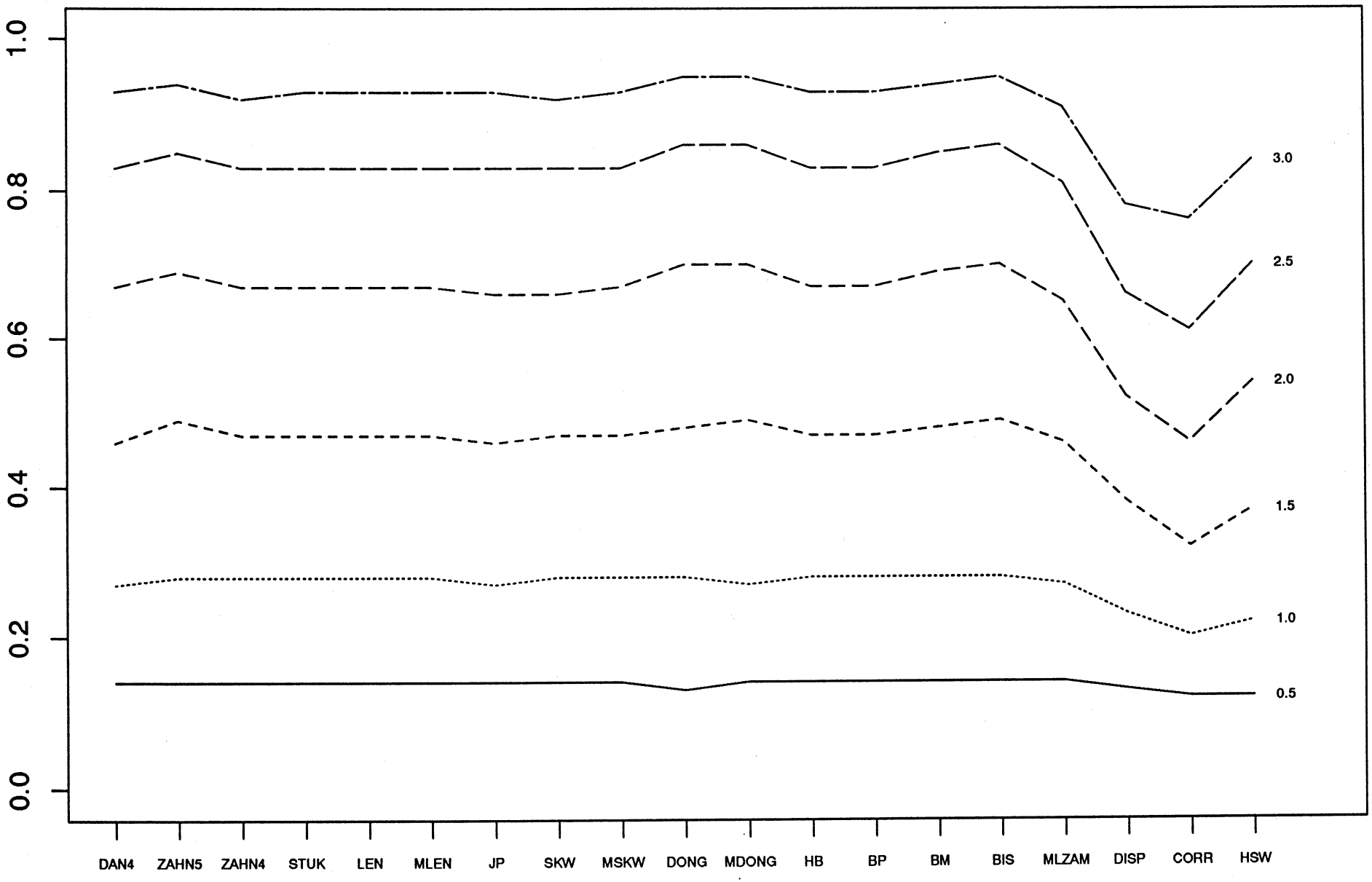

Figure 1b: Power for n=8
two active effects = .5(.5)3 sigma

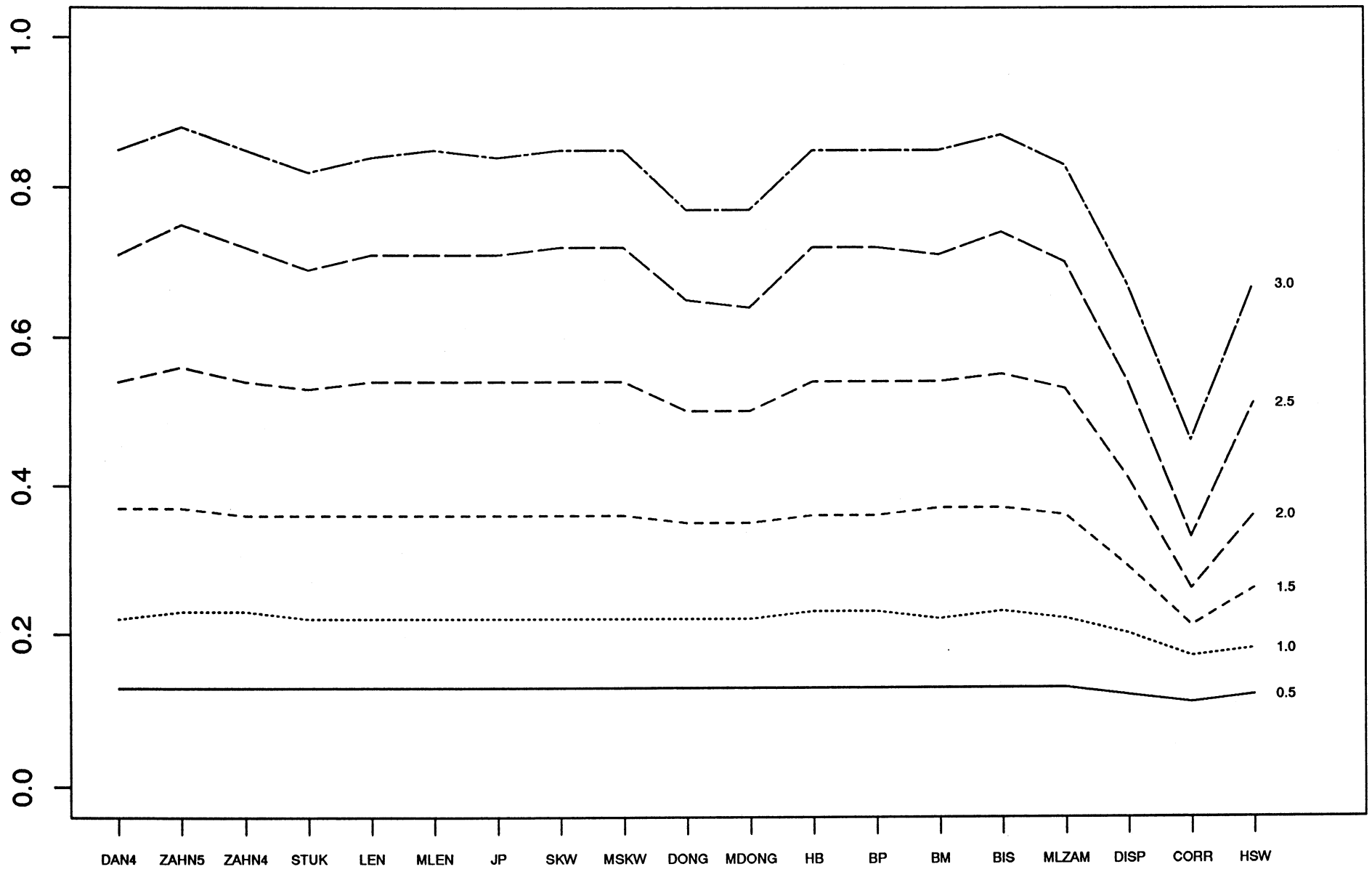

Figure 1c: Power for n=8
three active effects = .5(.5)3 sigma

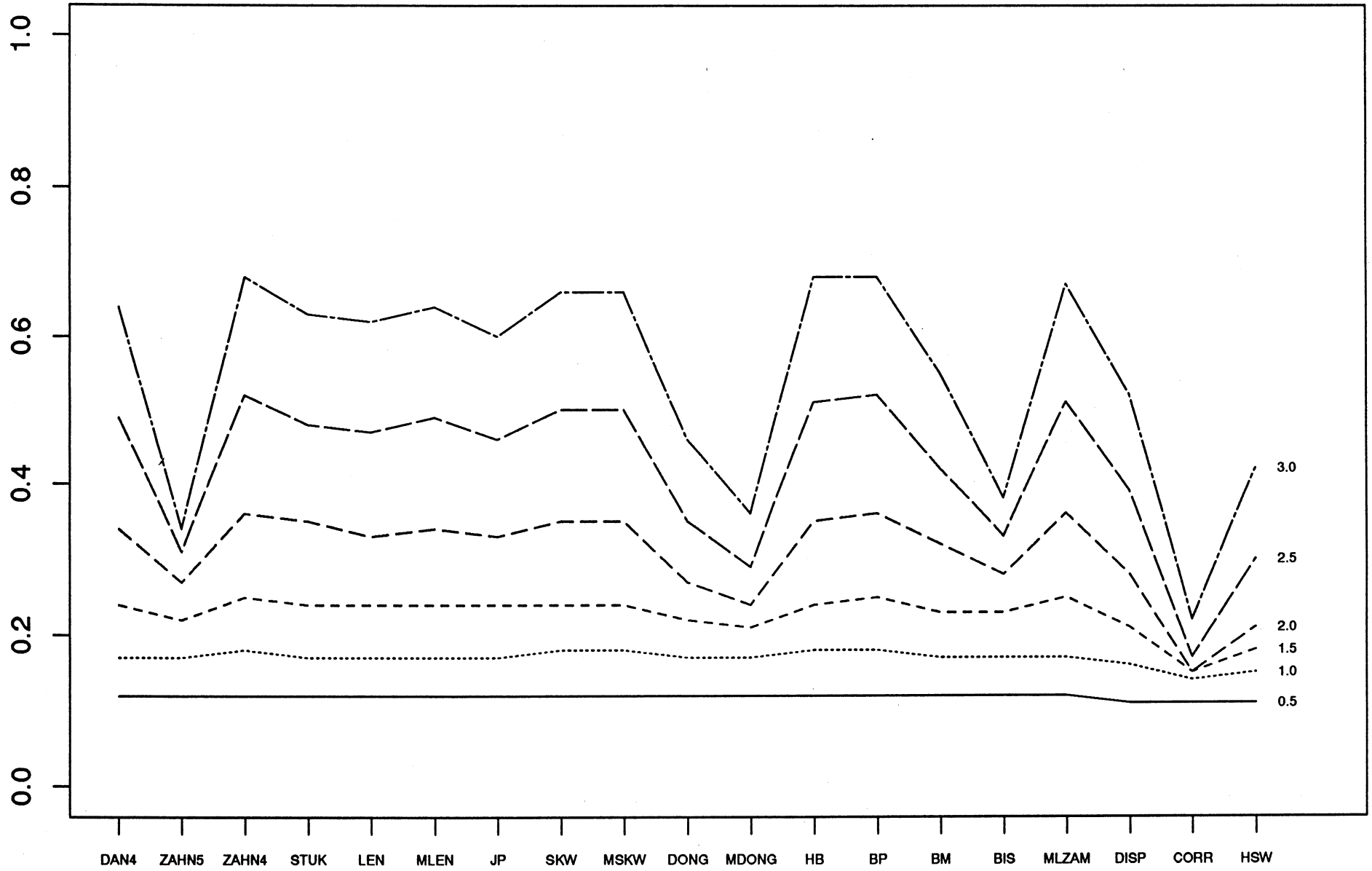

Figure 2a: Power for n=16
one active effect = .5(.5)3 sigma

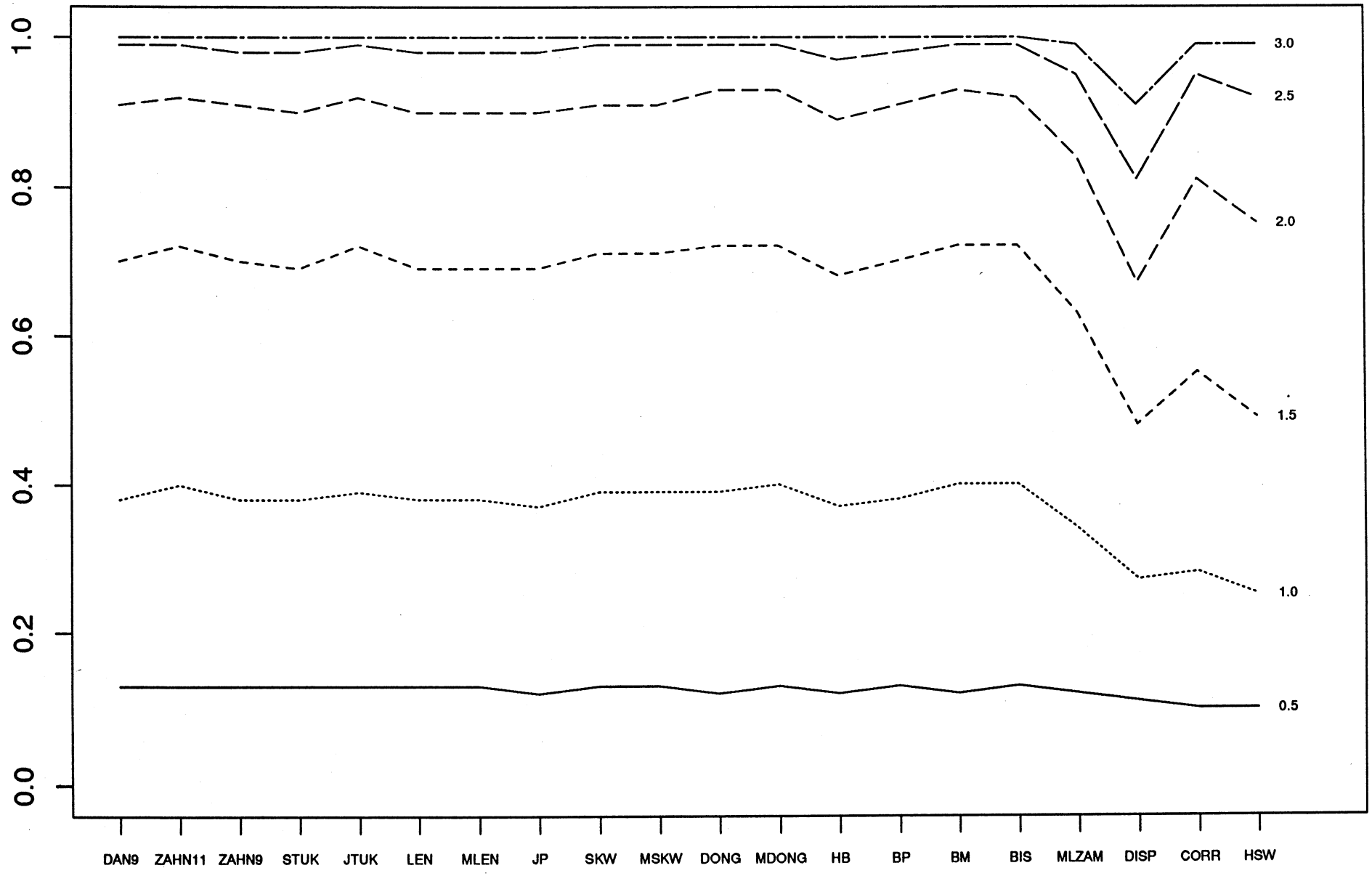

Figure 2b: Power for n=16
two active effects = .5(.5)3 sigma

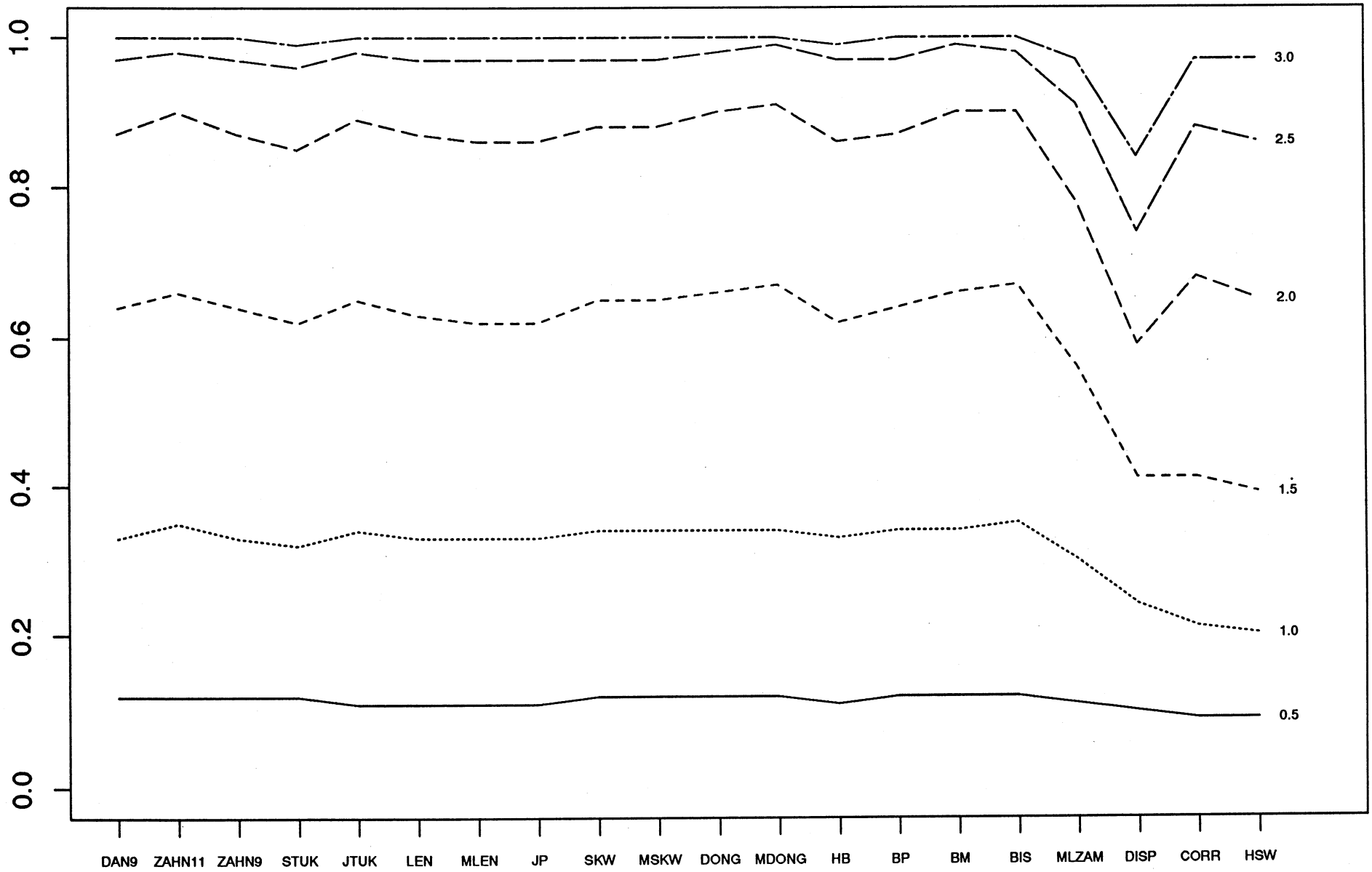

Figure 2c: Power for n=16
four active effects = .5(.5)3 sigma

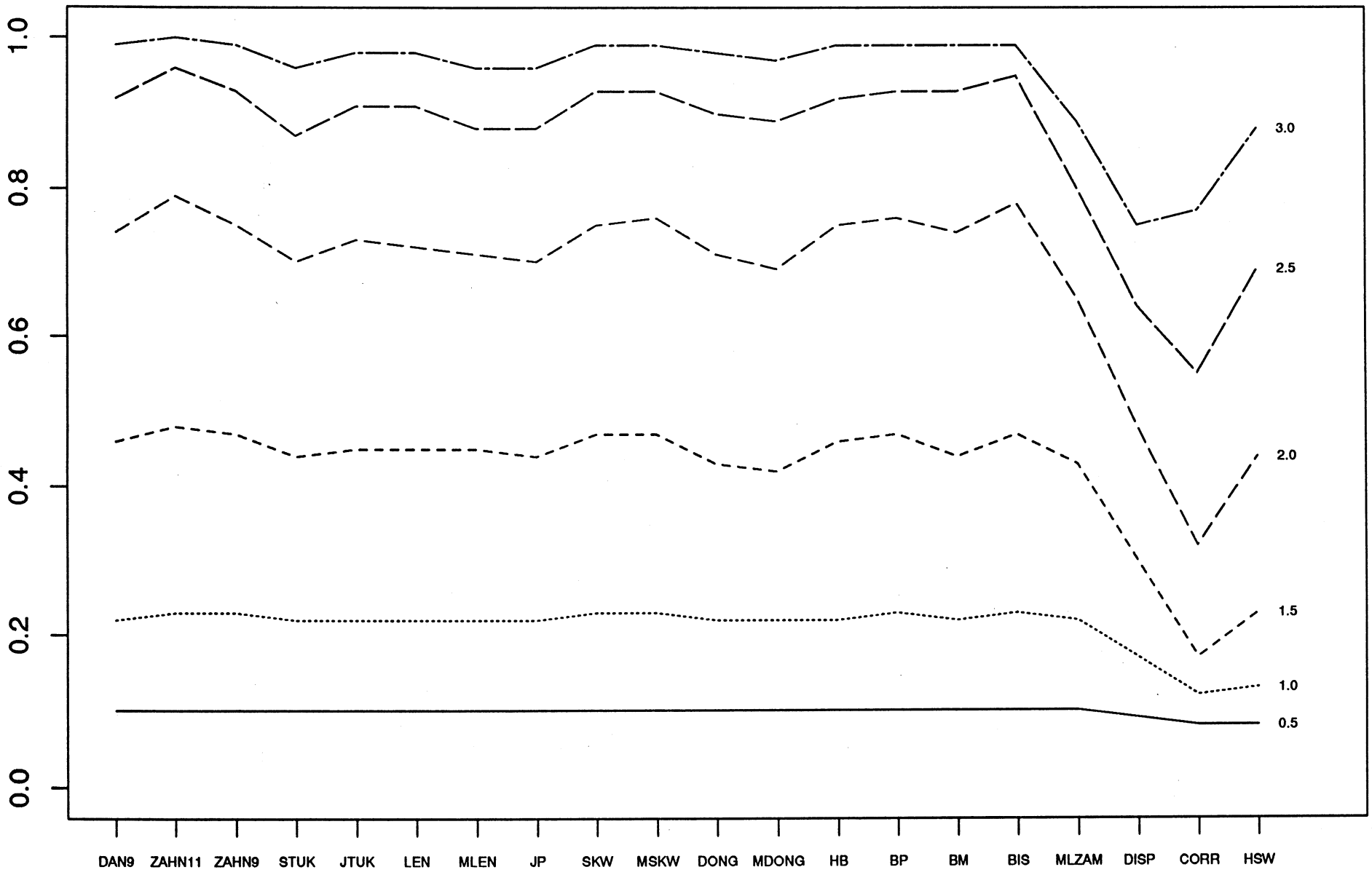

Figure 2d: Power for n=16
six active effects = .5(.5)3 sigma

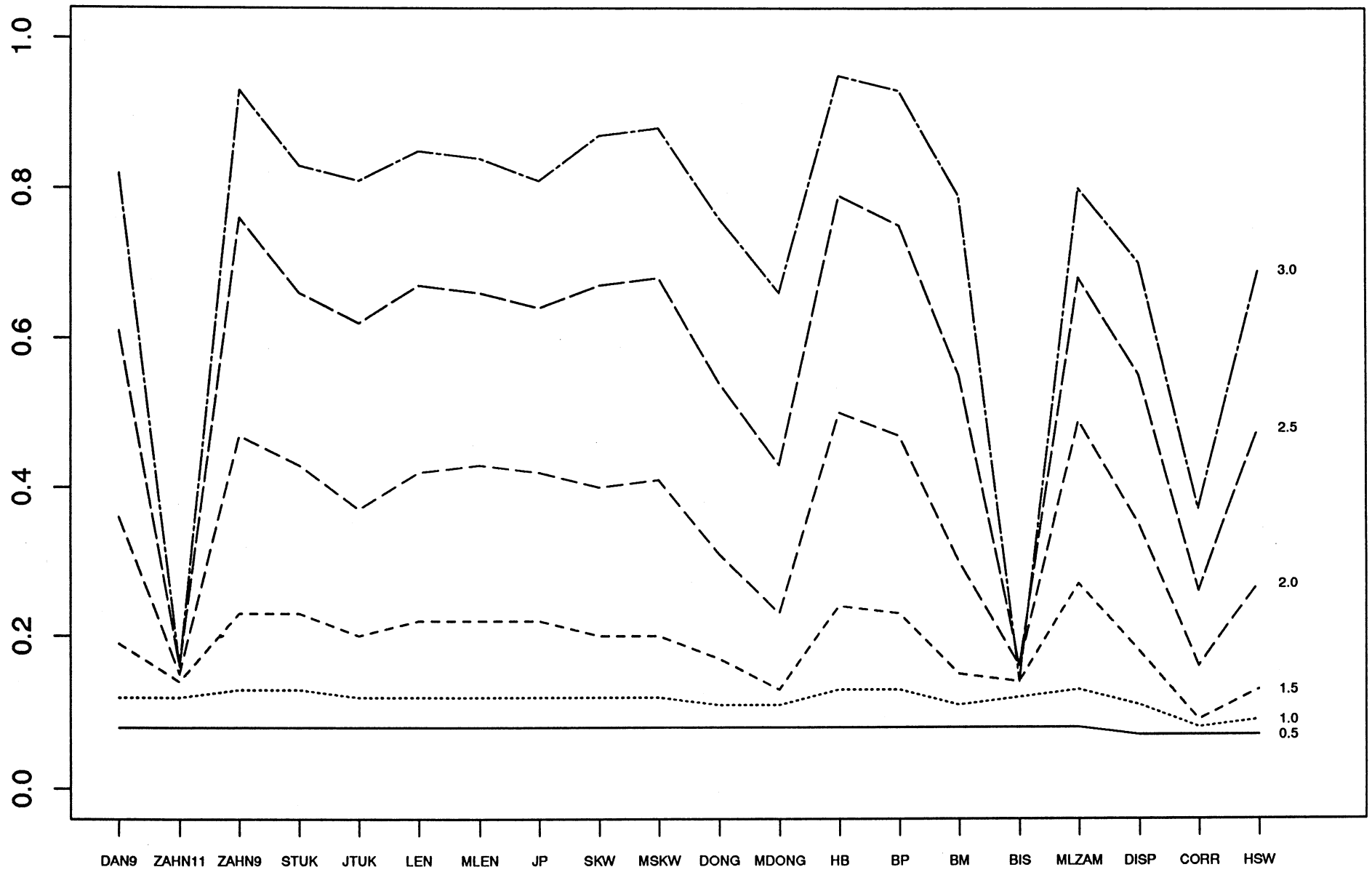

Figure 3a: IER for n=8
six inactive effects

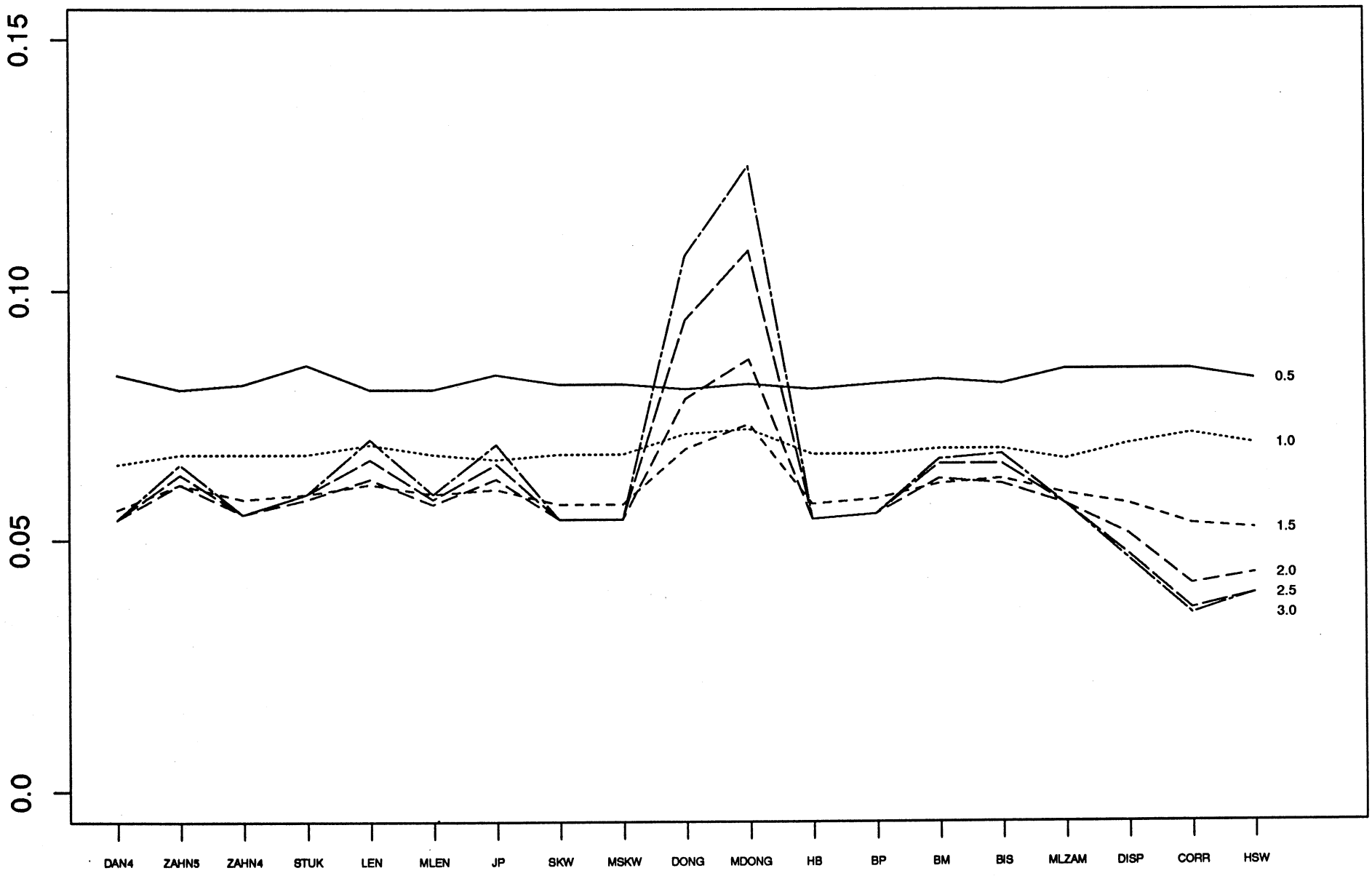

Figure 3b: IER for n=8
five inactive effects

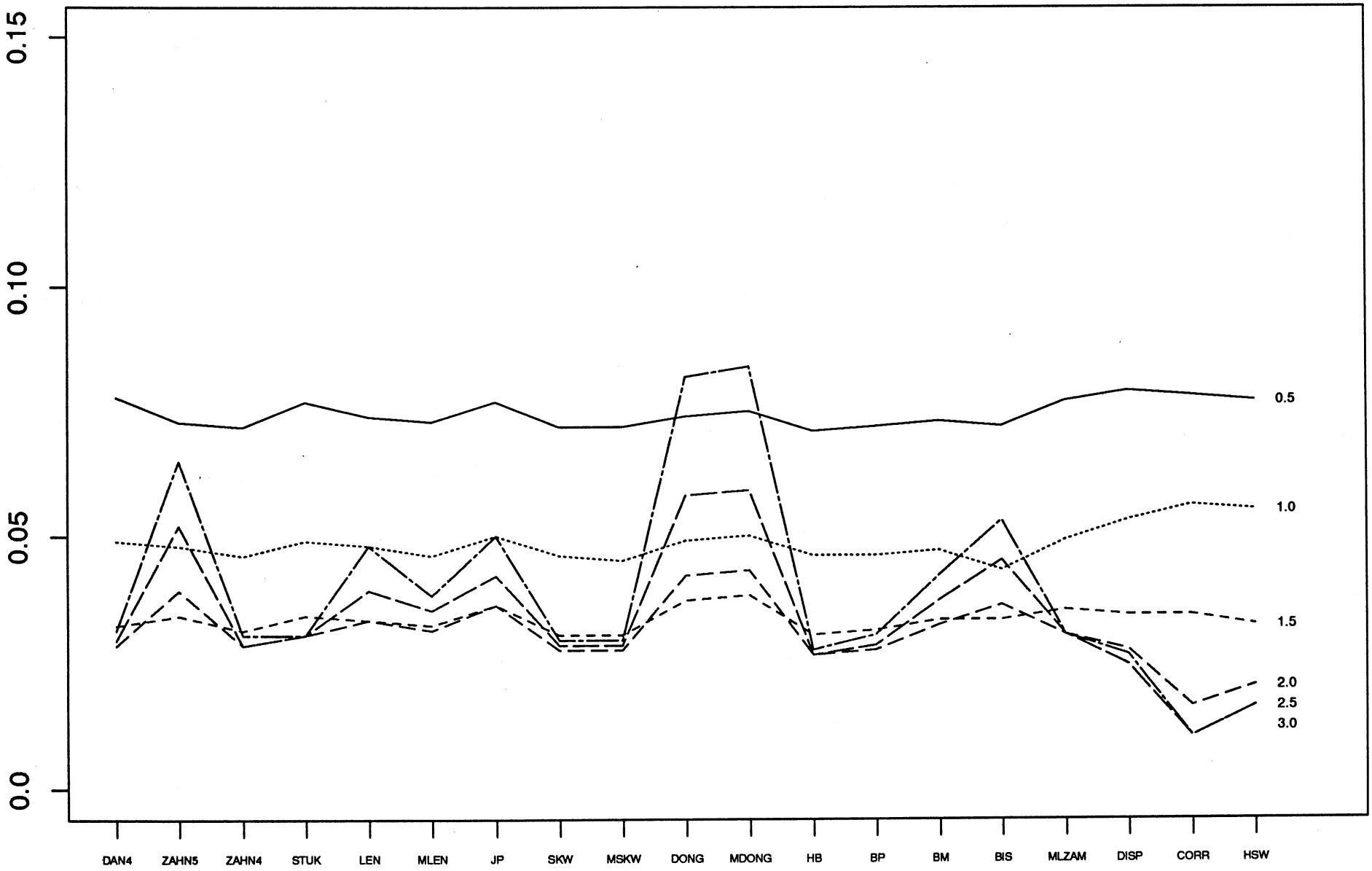

Figure 3c: IER for n=8
four inactive effects

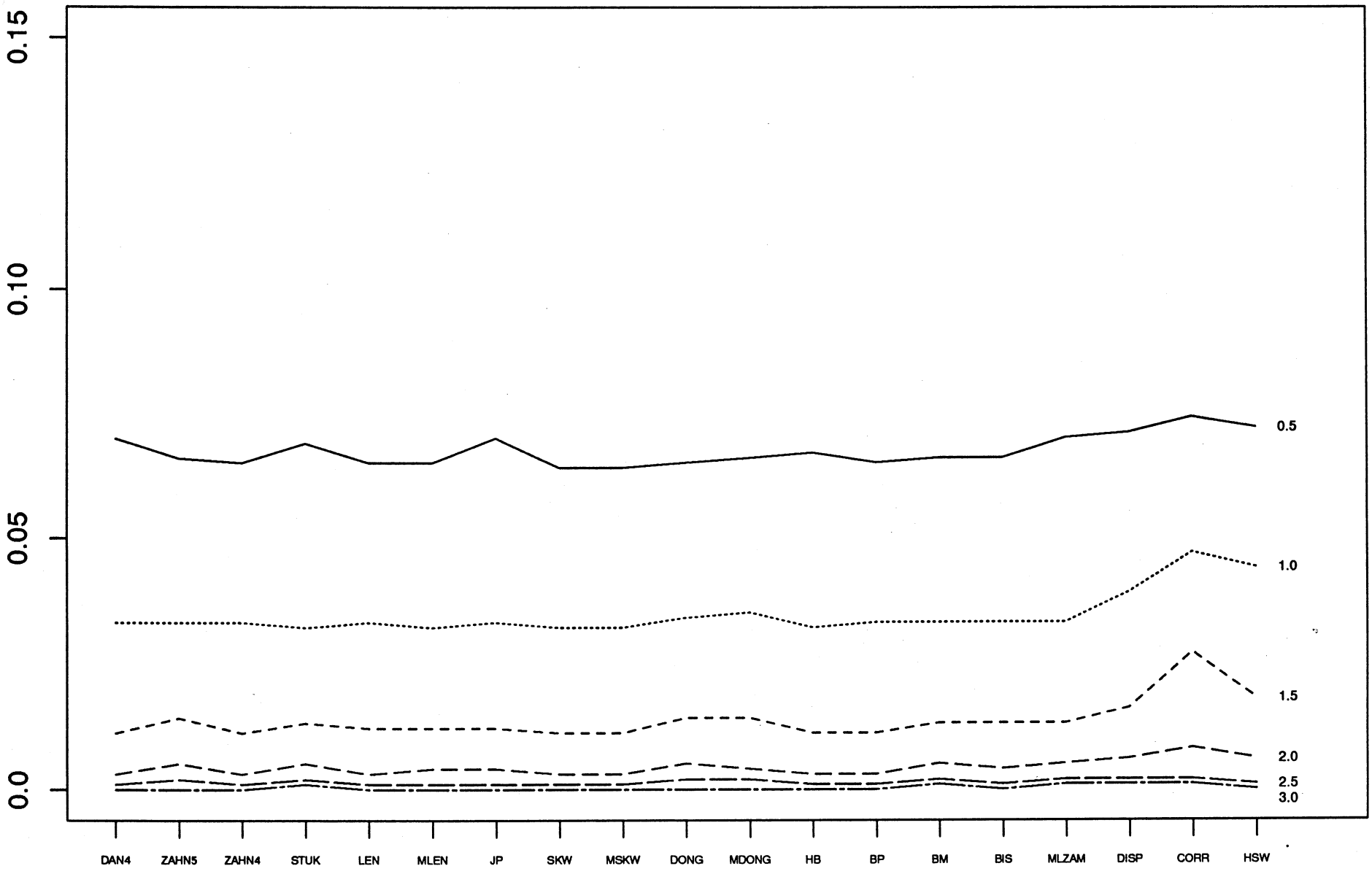

Figure 4a: IER for n=16
fourteen inactive effects

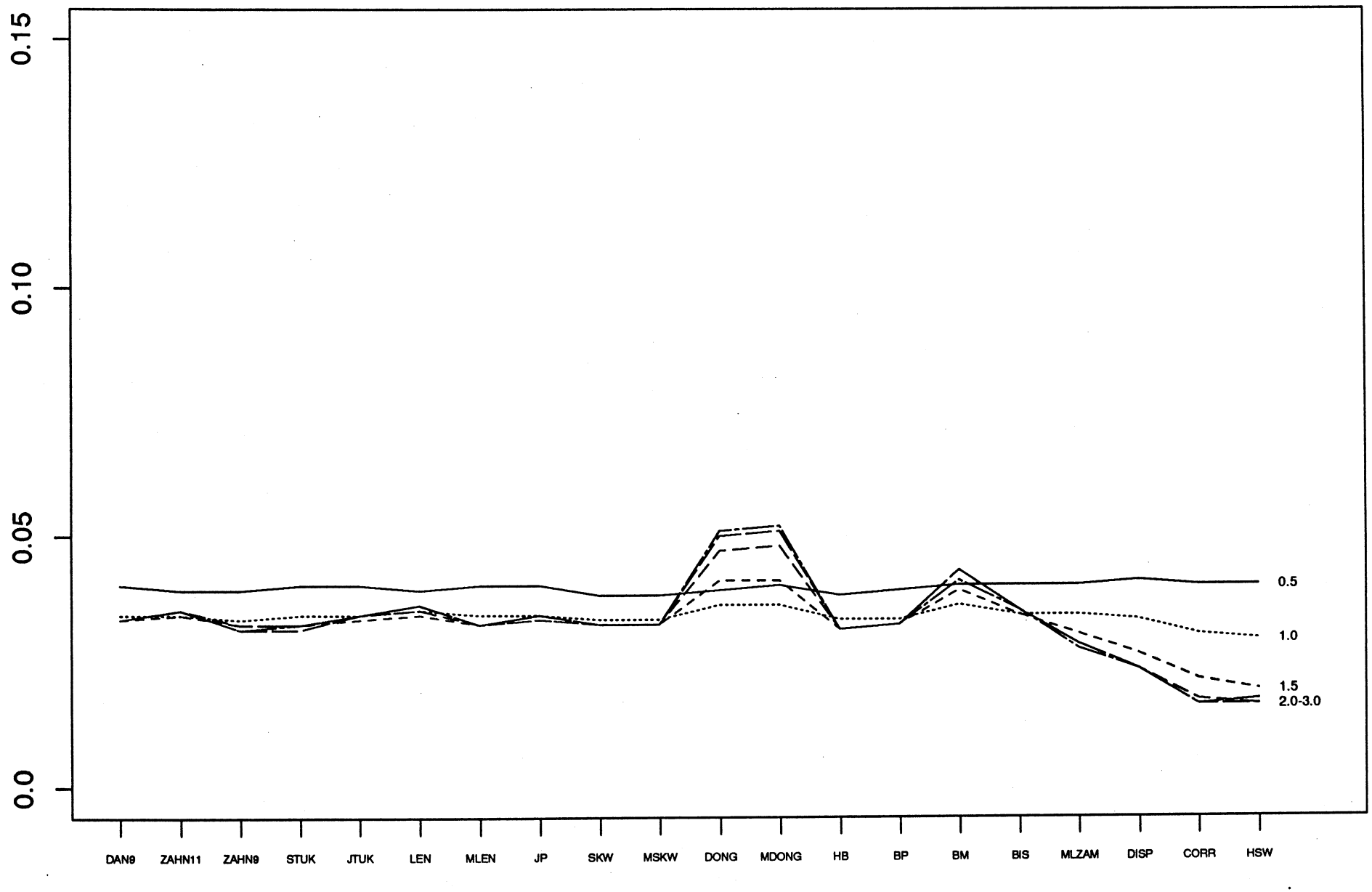

Figure 4b: IER for n=16
thirteen inactive effects

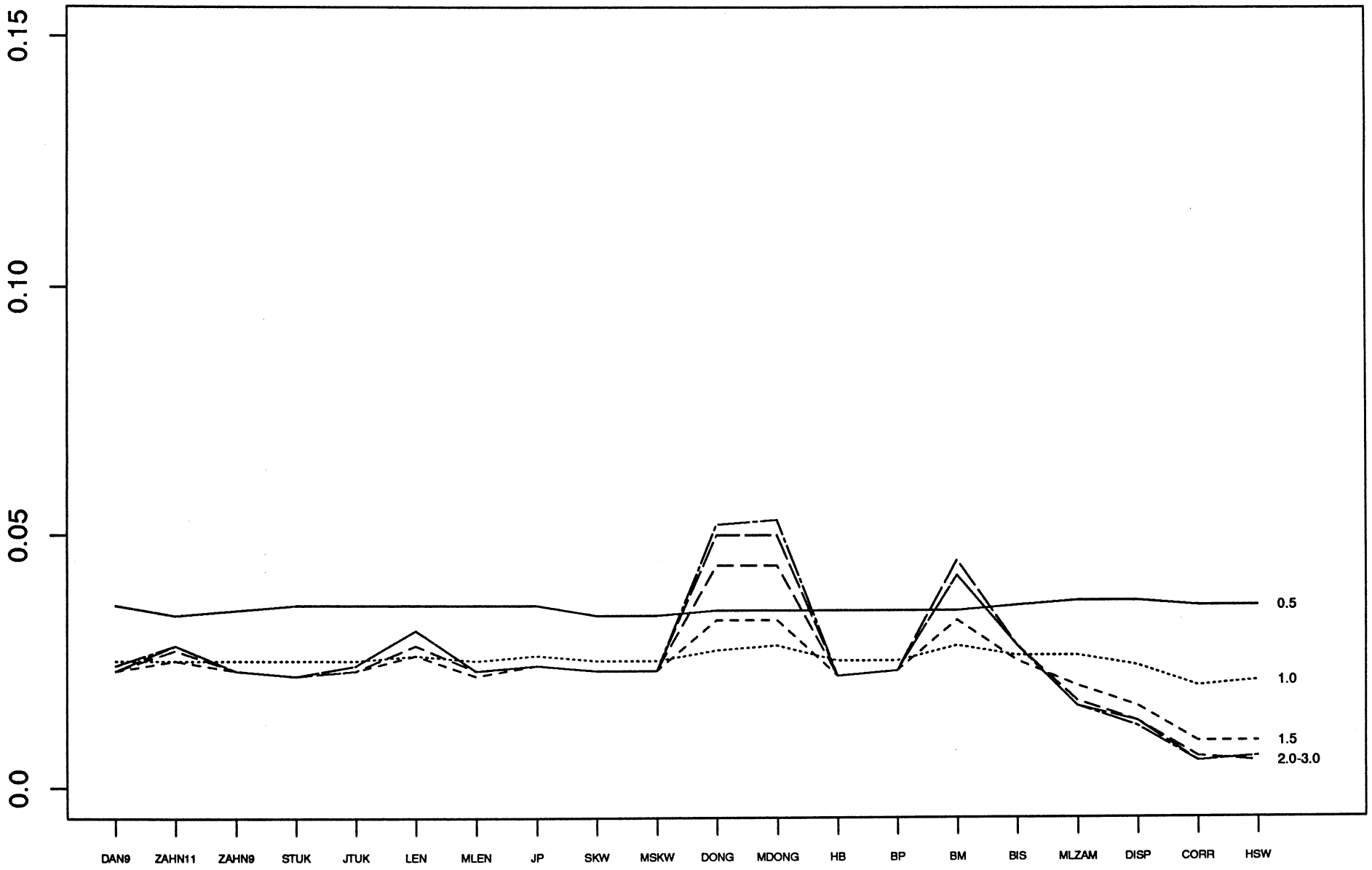

Figure 4c: IER for n=16
eleven inactive effects

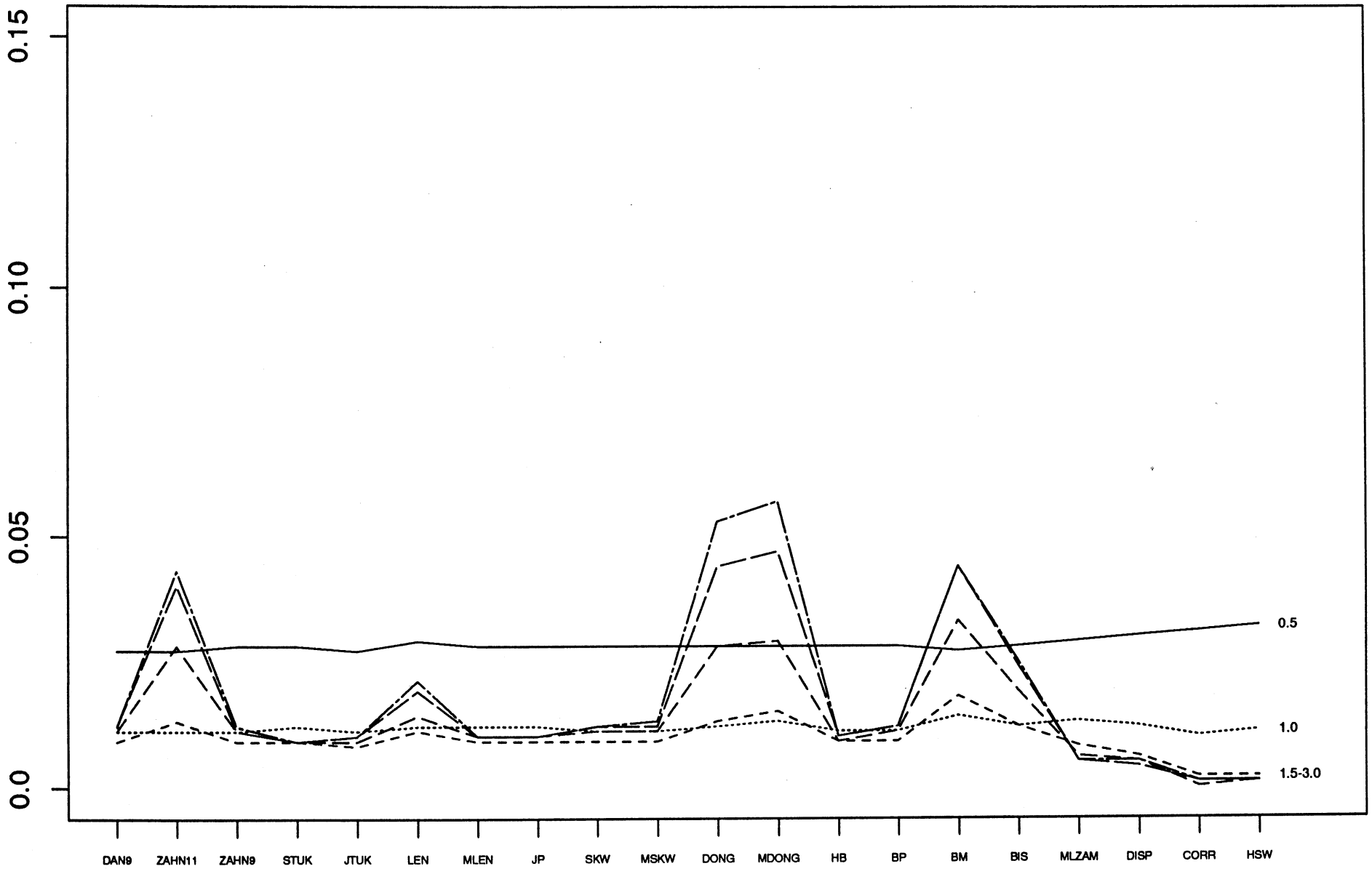

Figure 4d: IER for n=16
nine inactive effects