#2260

Inference for Effect Saturated Fractional Factorials

By Perry D. Haaland and Michael A. O'Connell*

Abstract

Industrial engineers and scientists often use experimental designs in which all degrees of freedom are used to estimate effects and consequently no classical estimate of the error is possible. Robust scale estimates provide an alternative measure of the error. In this study, several such scale estimators are evaluated based on the power of related significance tests. The pseudo standard error method (Lenth, 1989) provides the best overall performance. However, Lenth's t-approximation for critical values was found to be inaccurate so new tables are provided. Additional recommendations are made according to the experimenter's prior belief in the number of likely important factors.

KEY WORDS: orthogonal array, pseudo standard error, robust scale estimates, power, experimental design

1. Introduction

Due to cost and time restrictions, industrial experimentation is often geared toward the use of highly fractionated, unreplicated factorial designs. These designs typically allow no degrees of freedom for the estimation of error and are referred to by Box and Meyer (1986) as effect saturated designs. Because there is no independent estimate of the experimental error, identification of important effects lies outside the range of classical methods. This problem has motivated many publications including Daniel (1959), Zahn (1975a,b), Daniel (1976), Nelson (1982), Box and Meyer (1986), Lawson and Gold

*Perry D. Haaland is Research Fellow at Becton Dickinson Research Center, Research Triangle Park, NC 27709-2016 and Adjunct Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Michael A. O'Connell is Research Scientist II, Becton Dickinson Research Center.

(1988), Voss (1988), Wheeler (1988), Benski (1989), Haaland (1989), Lenth (1989), Stephenson, Hulting and Moore (1989), Berk and Picard (1991), Grize (1991), Stephenson (1991), Dong (1993), Le and Zamar (1993), and Schneider, Kasperski and Weissfeld (1993). In particular, Berk and Picard (1991) provide an excellent review of the literature and an extensive evaluation of several of the most popular methods for significance testing.

The objective here is to develop and examine methods of identifying nonnull effects in saturated fractional factorial designs. One statement of this problem is as follows (Zahn 1975b): Let β_i, i=1, 2, ... , k, be the true effects from a saturated fractional factorial design or orthogonal array. Suppose that the β_i are independent, normally distributed random variables with means μ_i, and common variance σ^2, respectively. Assume that some small number, relative to k, of the μ_i are nonzero and that there is no prior estimate of σ available. Then given the observed effects, $\hat{\beta}_i$,

(1) decide which, if any of the μ_i are nonzero and

(2) estimate σ.

An approach to this problem which motivates the use of robust estimators of scale is as follows: think of the estimated effects as a sample from a zero mean normal distribution (the null effects) contaminated by the nonnull effects. Use robust methods to find an estimate of the scale of the null effects that is insensitive to the nonnull effects. Then the estimated effects that are large compared to this scale estimate correspond to the nonnull effects.

As a first step towards solving this problem, Section 2 describes a family of robust scale estimators that can be used for significance testing. Special cases of this family include the pseudo standard error as proposed by Lenth (1989), a method referred to as the adaptive standard error that was proposed by Dong (1993), and the trimmed stan-

dard error which is equivalent to the trimmed ANOVA based method proposed by Berk and Picard (1991). A natural criterion for selecting the best scale estimator is that it should provide the most powerful test for identifying nonnull effects. In Section 3, construction of significance tests using robust scale estimators is described, and a table of critical values for the pseudo standard error based test is given. Section 4 reports the results of a simulation study carried out to evaluate the power of the various methods. In the first part of this study, the best values of certain tuning constants required to define the robust scale estimators are determined. In the second part of the simulation study, the scale estimators are compared and overall recommendations for use are made. Section 5 contains a discussion of the simulation results and compares the results to those obtained by Berk and Picard (1991) and Dong (1993). In Section 6, several examples are analyzed and discussed. Finally a short conclusion section is presented

2. A FAMILY OF ROBUST SCALE ESTIMATORS

The idea of robust scale estimation arises naturally in the analysis of unreplicated fractional factorials because the estimated effects can be thought of as a sample of zero mean normal deviates (the null effects) contaminated by the nonnull effects. If the scale of the null effects can be estimated, then the estimated effects that are large compared to this scale estimate correspond to the nonnull effects. Since it is not known *a priori* which effects are null, it is natural to want a test procedure that is insensitive to this lack of knowledge.

A brief overview of the literature on robust methods applied to significance testing in fractional factorial designs includes Zahn (1975 a,b) who advocated estimating the scale from a trimmed regression on the normal plot of effects. Lawson and Gold (1988) and Johnson and Tukey (1987) suggested robust regressions on the half-normal plot. Voss (1988) and Berk and Picard (1991) proposed using an ANOVA based estimate of scale after trimming some of the largest effects. Daniel (1959) and Lenth (1989) suggested

robust scale estimates based on order statistics after trimming the largest effects. Le and Zamar (1993) considered M-estimates of scale that downweight the largest effects. Dong (1993) proposed a method which combines elements of both Lenth's pseudo standard error and Berk and Picard's trimmed ANOVA. The family of scale estimators discussed in this section unifies the approaches taken by Daniel (1959), Lenth (1989), Berk and Picard (1991), and Dong (1993).

2.1 Definitions

For a two-level fractional factorial design, the estimated effect for a factor is the difference in averages between the two sets of n/2 observations at its high and low settings. However, the methods described here do not depend on how the effects are estimated and so "half effects" (coefficients from a regression model) could be used without loss of generality.

A family of robust scale estimators is now described. The corresponding estimates are each formed in two steps. First, an initial estimate of scale is obtained. The final scale estimate is then calculated using all estimated effects that are smaller than some multiple of the initial scale estimate. The form of the estimate calculated in the second step is chosen either for robustness or efficiency.

The initial scale robust estimate of scale, s_0, is calculated directly from $\hat{\beta}_1$, $\hat{\beta}_2$, ... $\hat{\beta}_k$, the estimated effects; namely

$$s_0(q) = a_0(q) \cdot \textit{quantile}\{q; |\hat{\beta}|_i, i=1,...,k\}$$

where $|\hat{\beta}|_i$, i=1,...k, are the absolute values of the estimated effects, *quantile*{q; X} is the q^{th} quantile of X, and $0 \le q \le 1$. Since the number of effects, k, is fixed in any application, for simplicity, the dependence on k is suppressed in this notation. The consistency constant, $a_0(q)$, is defined as

$$a_0(q) = \frac{1}{\Phi_0^{-1}(q)}$$

where $\Phi_0^{-1}(q) = \Phi^{-1}((q+1)/2)$ and Φ is the cumulative distribution function of the standard normal distribution. For example, when q=0.5, the initial scale estimator is defined as

$$s_0(0.5) = 1.4826 \cdot median(|\hat{\beta}_i|) .$$

The purpose of the initial scale estimate is to provide a basis for eliminating contaminating (nonnull) effects from the calculation of the final estimate of scale.

There are two approaches to calculating the final scale estimate. In the first, a robust scale estimate is calculated after trimming those effects which are large compared to s_0; for example, the pseudo standard error (PSE) (Lenth, 1989) estimator has the following form:

$$\hat{\sigma}_{PSE}(q,b) = a_{PSE}(q,b) \cdot median\{|\hat{\beta}_i| : (|\hat{\beta}_i| \le b \cdot s_0(q))\}$$

where $a_{PSE}(q,b)$ is a consistency constant, q is carried along from the initial scale estimator, and b is a second tuning constant. The PSE given by Lenth (1989) is equivalent (except for minor differences in the consistency constants) to $\hat{\sigma}_{PSE}(0.5, 2.5)$. The method proposed by Daniel (1959) is also very similar to the PSE. In particular, Daniel used q=0.683 in s_0 and also used this quantile instead of the median in the final scale estimator Note that $a_0(0.683)$ is approximately 1. He did not formally use a trimming threshold of $b \cdot s_0(q)$ but instead constructed the half-normal plot with slope $1/s_0(q)$ and trimmed large effects by inspection before recalculating the scale estimate. The consistency constant in the second stage was not corrected for the effects of trimming.

In the second approach, an efficient scale estimate is calculated after trimming those

effects which are large compared to s_0. This estimator, referred to herein as the adaptive standard error (ASE), is defined as follows:

$$\hat{\sigma}_{ASE}(q, b) = a_{ASE}(q, b) \cdot \sqrt{\frac{\sum_{|\hat{\beta}_i| \le b \cdot s_0(q)} \hat{\beta}_i^2}{m}}$$

where $a_{ASE}(q,b)$ is a consistency constant, m is the number of $|\hat{\beta}_i| \le b \cdot s_0(q)$, and b is a tuning constant. The estimator $\hat{\sigma}_{ASE}(0.5, 2.5)$ is equivalent (except for minor differences in the consistency constants) to an estimator proposed by Dong (1993). However, Dong also proposed that the method be iterated in order to improve the performance when there are more than a few nonnull effects. The impact of such iteration is discussed in Section 5.

Berk and Picard (1991) proposed a method based on trimmed ANOVA that is equivalent to a special case of the adaptive standard error. We call this estimator, the trimmed standard error (TSE). The TSE uses an efficient estimator of scale based on the effects remaining after trimming 100(1-α)% of the largest estimated effects; namely,

$$\hat{\sigma}_{TSE}(q) = a_{TSE}(q) \cdot \sqrt{\frac{\sum_{i=1}^{m(q)} |\hat{\beta}|_{(i)}^2}{m(q)}}$$

where $a_{TSE}(q)$ is a consistency constant, m(q) = *quantile*{q;1,...,k} = the number of effects included in the second stage, and $|\hat{\beta}|_{(1)} \le |\hat{\beta}|_{(2)} \le \ldots$ are the order statistics of the absolute values of the estimated effects. The estimator $\hat{\sigma}^2_{TSE}(0.6)$ can easily be seen to differ only by a constant from the mean square error (MSE) of the trimmed ANOVA proposed by Berk and Picard (1991). In particular,

$$MSE = \frac{n}{4} \cdot \frac{\hat{\sigma}^2_{TSE}(0.6)}{a^2_{TSE}(0.6)}$$

where n=k+1 (for effect saturated designs). Thus significance tests based on the TSE as described in Section 3 and the empirical tests proposed by Berk and Picard are equivalent. Note that the TSE can be seen to be a special case of the ASE as follows:

$$\hat{\sigma}^2_{TSE}(q) = \hat{\sigma}^2_{ASE}\left(q, \frac{1}{a_0(q)}\right);$$

that is, b is chosen to include the smallest 100*q% of the effects.

One additional estimator in this family that uses fixed trimming (as in the TSE) followed by the use of the median as in the PSE was considered. However, the properties of this estimator are not very good so no further discussion of it is included.

2.2 Estimation of consistency constants

As described above, the consistency constant used in the construction of s_0 can be determined directly from the sampling distribution of absolute values of a standard normal random variable. The consistency constant used in the second step can be determined empirically. In order to do this,10,000 samples of k effects from a standard normal distribution (μ=0, σ=1) were generated. Scale estimates were calculated for each sample. The consistency constant for an estimator was then calculated as the inverse of the mean of the scale estimates from the 10,000 samples for that estimator. Consistency constants are given in Table 1 for selected values of the tuning constants. Standard errors for the consistency constants are all less than 0.01. The tabulated values in Table 1 are used in the simulation study described in Section 4, but in practice the small differences between the Table 1 values and those used in the original methods can be expected to have only a minor influence on the results.

2.3 Discussion

The estimators described above can be expected to perform differently in the presence of varying numbers of nonnull effects. If the initial scale estimate provides a good basis for trimming nonnull effects, then the efficiency of the second stage estimator is its most important characteristic. On the other hand, if some nonnull effects contaminate the set of "null" effects used for the second stage estimate, then robustness of the second stage estimator is important.

Consider first the case in which there are only one or two large nonnull effects. In this case, it should be easy to find and trim them. The best estimator is then the one which trims the fewest null effects in the first stage and has the most efficient second stage. It is well known that the second stage estimator used in the ASE is efficient when there are no contaminating effects. Hence, the ASE should be the best scale estimator when there are few nonnull effects. It is also clear that as long as there are fewer than 100*(1-q)% nonnull effects, the TSE should be worse than the ASE because the TSE will trim too many null effects.

Now consider the case when there are many nonnull effects. For example, when k=15 and q=0.5, s_0 is a function of the 8th largest absolute effect (the median). If this effect becomes large, then so does the value of s_0. Thus, s_0 can "break down" when there are 8 nonnull effects. Following Hampel (1971) and Hoaglin, Mosteller and Tukey (1983), we define the breakdown bound of s_0 to be r/k where r is the greatest number of effects that can be replaced with arbitrary values while leaving s_0 bounded. Thus for k=15 and q=0.5, the breakdown bound of s_0 is 7/15. For each of the estimators described in this section, the breakdown bound is the same as that of s0.

As q approaches 0, s_0 has a higher breakdown bound and so provides more protection against contamination by the nonnull effects. Increased trimming (smaller b values)

also provides more protection. This protection is, however, achieved at the expense of efficiency. So a high breakdown bound is not desirable if there are few nonnull effects. For example, $\hat{\sigma}_{PSE}(0.45, 1.25)$ is more resistant to nonnull effects than $\hat{\sigma}_{PSE}(0.5, 2.5)$ at the expense of being less efficient. When there are many nonnull effects, there is also an increased chance that some of them will survive the trimming step. Then the robustness of the second stage is an important issue. Consequently, the PSE should generally perform better than the ASE for moderate to large numbers of nonnull effects.

3. SIGNIFICANCE TESTING

Once a robust scale estimate has been calculated, the important effects can be identified as those effects that are "large" in comparison to the scale estimate. A natural approach is to divide each effect by the scale estimate and compare the resulting statistics against critical values from a reference distribution. Thus, for example, a natural test based on the PSE would use the following test statistic:

$$t_{PSE} = \frac{|\hat{\beta}_i|}{\hat{\sigma}_{PSE}}.$$

The critical region of the test of H_0: $\mu_i=0$ is then $t_{PSE} > t_{PSE}(1-\alpha,k)$. The $100*(1-\alpha)\%$ critical level for k total effects, $t_{PSE}(1-\alpha,k)$, can be obtained empirically.

In this study, the calculation of critical values was based on 10,000 samples of k effects generated from the standard normal distribution ($\mu=0$, $\sigma=1$). For each sample, test statistics such as that given above for the PSE based test were calculated for each scale estimator. An empirical distribution function (edf) was then calculated for each test statistic using the absolute values of all k*10,000 effects. The critical values were obtained as quantiles of the edf.

Critical values for the PSE based test with values of the consistency constants as

originally given by Lenth are given in Table 2. (For the representative case, k=15, half-widths of 95% confidence intervals are ≤ 0.01 for critical values corresponding to $\alpha \leq 0.05$. The half-width is 0.05 for the critical value corresponding to $\alpha=0.01$. Critical values for the PSE based test as used in this paper may be obtained by using the consistency constants in Table 1 to adjust the values in Table 2.) Lenth (1989) recommended using critical values from the usual t-distribution with k/3 degrees of freedom. His approximation is based on comparing the empirical distribution of $\hat{\sigma}^2_{PSE}$ to chi-square distributions. The difference between the critical values obtained empirically as described above and the approximate values used by Lenth is great enough so that Lenth's approximation is not recommended for practical use. For example, t(.975,15/3)=2.571 whereas the corresponding empirical critical value is $t_{PSE}(.95,15)=2.153$. Consequently, using $\alpha=0.05$ in the PSE based test results in an empirical size of about 0.029 under Lenth's approximate procedure. Thus, Lenth's t-approximation results in an overly conservative test. Berk and Picard (1991) and Dong (1993) also noted this discrepancy.

Dong (1993) recommended critical values for the ASE based test using the usual t-distribution with degrees of freedom equal to the number of effects included in the final scale estimate. This results in slightly conservative values for the ASE; namely a 0.05 level t-approximation critical value results in an empirical size of about 0.042. This method works better for the iterated version of the ASE based test giving an empirical size of 0.053.

When Dong's method of determining degrees of freedom is applied to the PSE based test, the method works quite well for a 0.05 level test; namely, an empirical size of 0.051 is observed. However, the approximation is not as good at other α levels. For example at $\alpha=0.02$ the empirical size is 0.028 and at $\alpha=0.10$ the empirical size is 0.091. Thus, while determining degrees of freedom from the number of effects used in the final scale estimator is superior to Lenth's k/3 rule, it can not be generally recommended compared to

the use of the empirically determined critical values given in Table 2.

Berk and Picard recommended F-like tests from a trimmed ANOVA for use in the same context as described above. In Section 2, it is pointed out that this method is equivalent to using TSE based tests. Berk and Picard obtained critical values based on a numerical study similar to the one used here. These values agree closely with the corresponding critical values obtained in this study. For example, for k=15 and α=.05, Berk and Picard's critical value translates to 18.93 in comparison to the value of 18.97 obtained in this study.

4. NUMERICAL RESULTS

The performance of the different scale estimators within the framework described in the Section 3 is evaluated via a simulation study. For k=15, n=2000 sets of effects were generated for each of the cases described by the model given below. The study objective is to identify the scale estimator(s) which provides the most powerful test, and, as a result, make recommendations regarding the circumstances under which various methods should be used. Note that the case k=15 corresponds to 16 run effect saturated fractional factorial designs which are commonly used in industrial applications.

4.1 Description of the simulation study

Simulation studies are typically used to evaluate the performance of significance tests for effect saturated fractional factorials when some number, say r, of the k true means are set to values other than zero. For example, Daniel (1959) and Birnbaum (1959) considered the case in which r=1 of the means ranged from 1σ to 6σ. However, most studies include several values of r, and the values of nonzero means usually vary in a fixed pattern. For example, Zahn's (1975b) Type I model set r=1,2,4, and 6 and let the nonnull effects have equal mean values from 2σ to 8σ in increments of 2σ.

Berk and Picard (1991) described a more complicated study in which r=1,2,3,4, and 6

and for r=1, 4 and 6, the nonnull effects had equal means ranging from 1σ to 7σ in increments of 1σ. For r=2, the nonnull effects had means +/-fσ for f=1 to 7, and for r=3, the nonnull effects had means equal to $f\sigma/2$, $f\sigma$, and $3f\sigma/2$ for f=1 to 7. Zahn's (1975b) Type II scenario also had a complicated pattern of nonzero means. For r=2, he considered sets of nonzero means of the form $(2\sigma, 4\sigma)$, $(2\sigma, 6\sigma)$, $(2\sigma, 8\sigma)$, etc. For r=4, he considered nonzero means of the form $(2\sigma, 2\sigma, 4\sigma, 4\sigma)$, $(2\sigma, 2\sigma, 6\sigma, 6\sigma)$, etc.

In this study, a number of different models, all using k=15, including equal and unequal mean models, were considered. In general, results were readily interpretable within a model type, but mixing equal and unequal mean models for different values of r made the results appear less systematic. Only the results from the unequal mean model are presented here. In particular, values of r from 1 to 8 are considered, and for r=8, the means of the nonnull effects are set between $f\sigma/2$ and $3f\sigma/2$ for f=1-7. The pattern of the nonzero means is determined by multiplying the expected values of the order statistics from a standard normal distribution by a fixed value determined from the case r=8. The actual values of the means are shown in Table 3. Note that the spread of the nonnull effects is an increasing function of r.

4.2 Evaluation of tuning constants

The ASE and PSE methods each have two tuning constants; namely, q, the quantile determining which of the effects are used to estimate s_0, and b, the multiplier of s_0 used to trim large effects before calculating the final scale estimate. For simplicity, values of q are chosen so that *quantile*$\{q, |\hat{\beta}|_i, i=1,2, ..., k\} = |\hat{\beta}|_{(m)}$ where $|\hat{\beta}|_{(1)} \le |\hat{\beta}|_{(2)} \le ... \le |\hat{\beta}|_{(k)}$ are the order statistics of the absolute effects and m=6 to 10. Values of b=1.0 to 3.0 in increments of 0.25 are considered. In this scheme, for example, $\hat{\sigma}_{PSE}(0.5, 2.5)$ corresponds to m=8 (recall that k is fixed at 15). The TSE as proposed by Berk and Picard (1991), $\hat{\sigma}_{TSE}(0.6)$, corresponds to m=9.

The primary criterion for evaluating the scale estimators is the power to detect effects with nonzero means using the test described in Section 3. For this simulation study, the power is defined as the proportion of effects with nonzero means that are judged significantly different from zero at α=.05. The simulation results for a typical case, f=4 (that is, the nonnull effects have mean 4σ), are shown for the PSE test in Table 4, the ASE test in Table 5, and for the TSE test in Table 6. For $r \leq 3$, the standard errors in Tables 4-6 are all smaller than 0.01. For $r \geq 4$, the standard errors are all smaller than 0.006.

An examination of Tables 4-6 shows that the ASE based test generally has higher power for fewer effects, and that the PSE based test generally has higher power for greater numbers of effects, $r \geq 4$. The TSE based test in general is less powerful but the difference is fairly modest for fewer than 3 nonnull effects. The power of the ASE based test seems more sensitive to the choice of tuning constants, and the optimal tuning constants seem to be more dependent on f and r than for the PSE based test.

Now consider the problem of finding a single best set of tuning constants for the PSE test when the mean of the nonnull effects is f = 4 (Table 4). For r = 1-5 nonnull effects, the PSE with m = 8 and b = 2.5 is within about 0.01 of the highest power. For r = 7-8 nonnull effects, the PSE with m = 7 and b = 1.25 is within 0.01 of the highest power. For r = 6 nonnull effects, the best region is m = 8-9 and b about 1.5, and the two sets of values above (m=8, b=2.5 and m=7, b=1.25) are less than 0.04 from the highest power. This suggests that the PSE as defined by Lenth (1989) is generally a good compromise but that a more resistant estimator is useful when there are very many nonnull effects.

For the ASE based tests, m = 10 and b= 2.5 provides good power for r =1-3 nonnull effects. The values m = 8, b = 2.5 (as proposed by Dong, 1993) are almost as good for r = 1-3 nonnull effects and their power doesn't fall off as rapidly when there are more nonnull effects. However, there is no good choice for ASE based tests in comparison to the PSE based procedure when there are many nonnull effects. For the TSE based tests,

the power doesn't seem to be very sensitive to the choice of m, and the choice of m=9 as in Berk and Picard (1991) is a good compromise. For few nonnull effects it is nearly equivalent to the ASE and for many nonnull effects it is superior to the ASE. However, its power is generally less than that of the PSE based tests.

4.3 Recommended strategy

Figure1 shows a comparison of PSE and ASE tests based on the simulation results for the case f = 4 (recall that k = 15) with tuning constants (m,b) = (10,2.5) and (8,2.5) for the ASE and (m,b) = (8,2.5), and (7,1.25) for the PSE. The TSE based test with m=9 is also included for comparison. The ASE based test with m=8 and b=2.5 seems better than that with m=10, and b=2.5, providing higher power for more than 3 effects with non-zero means and about the same power for 3 or fewer effects with nonzero means. The PSE based test with m=8 and b=2.5 gives the best overall performance among all of the alternatives. The PSE based test with m=7 and b=1.25 provides the highest power when there are 7 or 8 nonnull effects. The TSE based test is not the best for any number of effects with nonzero means.

Before making a final recommendation, we consider the sizes of the various tests in the presence of effects with nonzero means. Based on the results of the simulation study, there is little difference among the tests in this regard. For example, there is a common maximum size of about 0.02 for all tests over all conditions. The reason for the small sizes in this case is that if there are nonnull effects, then either the corresponding estimated effects will be trimmed (in which case the estimate of scale is not affected) or they will be included in the scale estimate (in which case the scale estimate will be inflated). Thus the values of the test statistics for the null effects tend to be smaller because of the larger denominator. When these are compared to the critical values from the case in which all effects are null, they will be less likely to be judged significant. Hence the size of the test is actually reduced in the presence of nonnull effects.

Finally, consider the false positive rates or equivalently, in the terminology of Berk and Picard (1991), the "proportion of effects judged 'real' that are actually real". In general, the PSE based tests did not perform as well on this criteria as did the ASE or TSE based tests. For example, for the case, f=4 (mean of nonnull effects) with 4 nonnull effects (r=4), the proportions of effects judged 'real' that are actually real are 0.96, 0.96, 0.98, 0.98, and 0.99 for the PSE (m=7, b=1.25), PSE (m=8, b=2.5), ASE (m=8, b=2.5), ASE (m=10, b=2.5), and TSE (m=9), respectively. For 3 and 5 nonnull effects, the order is the same and the values range from 0.93 to 0.97 and from 0.98 to 1.00, respectively. For comparison, the corresponding maximum differences in power among these tests for f=4 and r=3-5 are 0.15, 0.10, and 0.15, respectively. Hence the benefits of increases in power are not in general greatly discounted by corresponding differences in rates of false positives.

These results held in general for all values of f evaluated in the simulation study (in the interest of conserving space, these results are not shown). Therefore, the following recommendations can be made for the case k=15:

- If only one estimator can be used, the PSE based test as defined by Lenth (1989) (q=.5, b=2.5) consistently provides high overall power except in the case of 7 or 8 effects with nonzero means (r=7 is the natural breakdown bound for PSE when k=15).

- If there is *a priori* reason to suspect that there are few effects, r=1-3, the ASE based test with q=.5 and b=2.5 as defined by Dong (1993) is recommended. This corresponds roughly to the prior $\alpha \leq 0.2$ in Box and Meyer's (1986) notation where α is the expected proportion of effects with nonzero means (3/15 = 0.2). In practical terms, this case corresponds to early screening experiments where there are many factors and only a few are thought to be significant. The ASE based test should not be used routinely except in this case.

- If there is *a priori* reason to suspect that there are a moderate number of effects, r=4-6, the PSE based tests with q=.5 and b=2.5 is recommended. This corresponds roughly to the prior $0.2<\alpha\leq 0.4$ which would be relevant to intermediate screening experiments where there are fewer factors and there are likely to be important interactions.

- If there is *a priori* reason to suspect that there are a large number of effects, r=7-8, the modified PSE based test with (m=7,b=1.25) is recommended. This corresponds roughly to the prior $\alpha>0.4$ which is relevant to late screening experiments where there are few factors, most of which are thought to be important and there are important interactions.

- If the experimenter strongly prefers to use an ANOVA program, then the trimmed ANOVA procedure as described by Berk and Picard (1991) which is equivalent to the TSE is an acceptable alternative as long as there aren't too many effects with non-zero means; say, $r \leq 5$.

5. DISCUSSION

The differences reported in this study are generally modest but systematic. For example, the power of the TSE based test procedure (m=9) differs from that of the recommended test procedure by 0.05 or less for $r\leq 6$ and by 0.14 and 0.21 for r=7 and 8, respectively. For $r\leq 2$, the ASE has an advantage in power compared to the PSE of less than 0.05. For r=3, they are equivalent. For $4\leq r\leq 8$, the PSE advantage over the ASE is between 0.03 and 0.12. Thus for few numbers of nonnull effects, say $r\leq 3$, the choice of method doesn't have very much impact. However, one can expect systematic improvements in power using the recommendations given above for larger numbers of nonnull effects.

We believe that the all around good performance of the PSE with q=.5 and b=2.5 is

likely to hold for other sample sizes. However, the recommendation for large numbers of nonnull effects is a more difficult question. The highly resistant version of the PSE could be viewed as having q = 7/15 = 0.467. It is not immediately clear that q = 0.467 would necessarily be the best choice over a wide range of k. This question remains open for further study.

Berk and Picard (1991) carried out an extensive simulation study of the following five methods: an ANOVA based pooling procedure due to Voss (1988), a testing procedure using the PSE as described by Lenth (1989), a Bayesian approach proposed by Box and Meyer (1986), Zahn's (1975a,b) trimmed regression, and their 40% trimmed ANOVA procedure. They found no consistent differences in power among the different test methods. There are a number of possible reasons why Berk and Picard's study didn't lead them to the same conclusions as reached here. Probably most important is that in this study each of the methods is adjusted to have the same empirical size in the null case whereas Berk and Picard repeatedly adjust their proposed test (the TSE) to have the same empirical size as the one they are comparing against. For example, Lenth's test as originally proposed has empirical size of 0.03 so Berk and Picard adjusted their test to have the same size and called it "60%(Lenth)". Thus the comparison of Lenth's method to the 60% method would be fair, but the series of 2-way comparisons with other tests makes it harder to determine which method is really the best overall. We also found early on that when the simulation model mixed equal and unequal mean models for various values of r (as in Berk and Picard's study) the results are not as clear. Finally, Berk and Picard didn't consider larger numbers nonnull effects (r=5, 7 and 8). We found the PSE based test to have the best performance in this case.

Dong (1993) conducted a numerical study comparing the PSE and ASE based test procedures and concluded that the ASE based test performed noticeably better than the PSE based test. In particular, Dong says that "For higher percentages of active contrasts

(e.g. 20%, 40%), Lenth's method is much better than the initial method [based on s_0] and the proposed method [iterated ASE] is better than Lenth's [PSE]." Since this conclusion directly contradicts the results of this paper, we repeated part of Dong's study in order to further investigate his claims. We found that when the methods are compared at the empirical size level of 0.05 rather than the much more conservative size considered by Dong, the PSE based test procedure dominates the ASE test and both the iterated version of the PSE and the iterated version of the ASE. Details of the study are available from the authors.

6. EXAMPLES

To illustrate the relative performance of the significance tests recommended in the previous section, three examples are analyzed. In general, the examples point out that (i) when there are few effects all of the methods are essentially the same and (ii) when there are many effects, the ASE and TSE methods are not as powerful as the PSE methods.

Box and Meyer (1986) discuss four 16-run experiments with all factors at two levels. We refer to these as B&M-I, B&M-II, B&M-III and B&M-IV and give Pareto plots (Haaland, 1989) and half-normal plots for B&M-1 and B&M-IV in Figures 2 and 3. B&M-II and B&M-III aren't discussed here because all of the methods give the same results and are in accord with the conclusions of Box and Meyer.The graphical analysis of B&M-I and B&M-IV (Figures 2 and 3) features vertical lines representing the cut-offs at which effects are judged to be significantly different from zero based on the four estimators of scale described above. These are $\hat{\sigma}_{PSE}(0.5, 2.5)$, $\hat{\sigma}_{PSE}(0.467, 1.25)$, $\hat{\sigma}_{ASE}(0.5, 2.5)$ and $\sigma_{TSE}(0.6)$. In the following discussion, the associated tests are referred to as PSE(0.5,2.5), PSE(0.467,1.25), ASE and TSE. In the notation of the simulation study, q=0.467 corresponds to m=7, q=0.5 corresponds to m=8, and q=0.6 corresponds to m=9.

In B&M-I (Figure 2), test procedures based on both PSE and TSE estimators of scale declare 3 effects to be significantly different from zero. This agrees with posterior probability plot conclusions of Box and Meyer (1986a). The ASE based test identifies 4 important effects. The Box and Meyer analysis does not produce a posterior probability of greater than 0.5 for the 4th effect for any values of the priors $0.1<\alpha<0.3$ and $5<k<15$. The ASE based test in this case may thus be too liberal in the detection of effects.

In B&M-IV (Figure 3) the PSE(0.467,1.25) based test declares 5 effects to be significantly different from zero, the TSE based test 2 effects, the PSE(0.5,2.5) based test 2 effects and the ASE based test 1 effect. The Box and Meyer analysis doesn't identify any nonnull effects for $\alpha=0.2$, $k=10$, however in the ranges of $0.1<\alpha<0.3$ and $5<k<15$, 8 different effects have posterior probabilities greater than 0.5. The half-normal plot (Figure 3) suggests that there may be three groups of effects; 2 large effects, 7 moderate effects, and 6 negligible effects. Not surprisingly, none of the procedures identifies all 9 of the largest effects although the PSE(0.467,1.25) does suggest that 5 of these effects may be nonnull.

The last example concerns the optimization of porous-carbon electrode structure (Kannan et al. 1988). The Pareto plot and half-normal plot of effects are given in Figure 4. The half-normal plot of effects suggests that there are 3, 5 or 7 nonnull effects. The PSE(0.5,2.5) and PSE (0.467,1.25) based tests each declare 5 effects significant. The ASE and TSE based tests find 3 effects. In comparison, the Box and Meyer procedure (Figure 5) also finds 3 nonnull effects for $\alpha=0.2$ and $k=10$, and up to 7 nonnull effects for $0.1<\alpha<0.3$ and $5<k<15$.

The examples B&M-I, B&M-II, and B&M-III give practical evidence that all of the test methods perform essentially the same when there are few nonnull effects. The Kannan et al. (1988) example illustrates the general power of the PSE(0.5,2.5) based test when there are more effects (possibly 5). The B&M-IV example is evidence of the difficulty in

identifying nonnull effects when there are potentially many nonnull effects and supports the conclusion that the PSE(0.467,1.25) based test is the most effective in this case.

7. CONCLUSIONS

The test for nonnull effects based on the pseudo standard error (PSE), as proposed by Lenth (1989), has good properties over a broad range of numbers and sizes of effects. This conclusion is based on an extensive evaluation of the power of significance tests to identify nonnull effects for 16 run designs. Critical values are given in Table 2 for use with this procedure. At least for the case of 16 run designs, critical values for the PSE based test at the α=0.05 level can also be generated from a t distribution with degrees of freedom equal to the number of effects used in the final estimate. However, this relationship has not been verified for other sample sizes. A more resistant version of the PSE based test, obtained by using different tuning constants, is recommended for special situations in which there are many nonnull effects. The trimmed standard error as proposed by Berk and Picard (1991) is satisfactory when there are not too many effects and there is a strong motivation to use an ANOVA based approach. The adaptive standard error as proposed by Dong (1993) is generally appropriate to use only when there is *a priori* reason to believe that there are not many nonnull effects. Iteration of either the PSE or ASE estimating procedure doesn't improve performance of the associated tests as long as critical levels are chosen to give a test with empirical size of about 0.05.

ACKNOWLEDGMENTS

We wish to thank Pat Burns, Doug Martin and Deborah Donnell of Statistical Sciences, Inc., Seattle, WA, Peter Bloomfield of North Carolina State University, Bruce Belanger of the Becton Dickinson Research Center and North Carolina State University, and Chris Triggs of the University of Auckland, New Zealand for helpful comments and assistance. The author's would also like to acknowledge the helpful comments of two anonymous referees and an associate editor.

REFERENCES

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 with replication." *Technometrics* **1**, 342-357.

Box, G.E.P. and Meyer, R.D. (1986). "An Analysis for Unreplicated Fractional Factorials." *Technometrics* **28**, 1-18.

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: Wiley.

Dong, F. (1993). "On the Identification of Active Contrasts in Unreplicated Fractional Factorials." *Statistica Sinica* **3**, 209-217.

Grize, Y. L. (1991). "Plotting Scaled Effects from Unreplicated Orthogonal Experiments." *Journal of Quality Technology* **23**, 205-212.

Haaland, P.D. (1989). *Experimental Design in Biotechnology.* Marcel Dekker, NY.

Hampel, F.R. (1981). "A General Qualitative Definition of Robustness." *Annals of Mathematics and Statistics*, **42**, 1887-1896.

Hoaglin, D.D., Mosteller, F., Tukey J.W. (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, NY.

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*, C. L. Mallows (Ed), New York: Wiley, 171-244.

Kannan, A.M., Shukla, A.K., and Hamnet A. (1988). "Fractional-factorial Design of a Porous-Carbon Fuel-Cell Electrode." *Journal of Applied Electrochemistry* ***18***, 149-153.

Lawson, J.S. and Gold, L. (1988). "Robust Estimation Techniques for Use in Analysis of Unreplicated d^k and d^{k-p} Designs." unpublished manuscript delivered at 1988 Annual Meetings of the American Statistical Association.

Le, N.D. and Zamar, R. (1993 in press). "An Optimal Test for Effects in 2^k Factorial Design Without Replicates." *Journal of Statistical Computation and Simulation.*

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

Nelson, L. S. (1982). "Analysis of Two-Level Factorial Experiments." *Journal of Quality Technology* **14**, 95-98.

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**, 18-17.

Stephenson, W. R. (1991). "A Computer Program for the Quick and Easy Analysis of Unreplicated Factorials." *Journal of Quality Technology* **23**, pp. 63-67.

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.

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.

Wheeler, D.J. (1988). *Understanding Industrial Experimentation.* Statistical Process Controls, Knoxville, TN.

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.

/stats/home/pdh/papers/graff/robust/inference.ncsu.doc. .

Table 1: Consistency Constants for Selected Estimators for Various Numbers of Effects (k) Based on 10,000 Simulated Samples

k	$a_{ASE}(.5,2.5)$	$a_{PSE}(.5,2.5)$	$a_{TSE}(.6)$
7	1.13	1.45	2.08
11	1.10	1.47	2.24
15	1.09	1.48	2.07
17	1.09	1.48	2.12
23	1.08	1.49	2.23
31	1.07	1.49	2.19

Table 2. Revised Critical Values for the PSE Based Test as Defined by Lenth (1989) for Various Numbers of Effects (k) Based on 10,000 Simulated Samples

k	α=.20	α=.15	α=.10	α=.05	α=.01
7	1.21	1.41	1.71	2.31	5.17
11	1.25	1.44	1.71	2.20	4.14
15	1.26	1.44	1.70	2.15	3.67
17	1.26	1.45	1.70	2.14	3.50
23	1.27	1.44	1.69	2.09	3.22
31	1.27	1.44	1.68	2.06	3.05

Table 3. Simulation Model for Nonzero Means

Number of nonzero means (r)	Model for nonzero means (overall average is f, σ=1)
1	f ×
2	f × (0.795, 1.205)
3	f × (0.697, 1.000, 1.303)
4	f × (0.634, 0.896, 1.104, 1.366)
5	f × (0.589, 0.827, 1.000, 1.173, 1.411)
6	f × (0.553, 0.776, 0.930, 1.070, 1.224, 1.447)
7	f × (0.524, 0.736, 0.877, 1.000, 1.123, 1.264, 1.476)
8	f × (0.500, 0.703, 0.835, 0.947, 1.053, 1.165, 1.297, 1.500)

Table 4. Simulation Results: Power of PSE Based Test for Various Values of the Tuning Constants m = the Number of Effects Included in the Initial Scale Estimator and b = the Multiplier Used in Trimming Effects before the Second Scale Estimator

m	b								
	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
r=1 effect with nonzero mean									
6	0.695	0.728	0.764	0.801	0.839	0.873	0.880	0.888	0.896
7	0.738	0.771	0.813	0.850	0.877	0.888	0.893	0.900	0.904
8	0.788	0.801	0.840	0.874	0.892	0.896	0.904	0.908	0.910
9	0.829	0.832	0.859	0.886	0.901	0.906	0.907	0.909	0.912
10	0.858	0.866	0.880	0.898	0.910	0.914	0.912	0.912	0.858
r=2 effects with nonzero means									
6	0.639	0.670	0.702	0.735	0.764	0.790	0.808	0.822	0.831
7	0.674	0.704	0.735	0.770	0.799	0.808	0.820	0.834	0.840
8	0.721	0.736	0.772	0.804	0.824	0.838	0.845	0.850	0.852
9	0.747	0.759	0.788	0.818	0.834	0.846	0.852	0.856	0.854
10	0.780	0.787	0.808	0.835	0.848	0.855	0.857	0.858	0.780
r=3 effects with nonzero means									
6	0.586	0.624	0.661	0.692	0.723	0.740	0.757	0.769	0.773
7	0.628	0.663	0.700	0.733	0.752	0.755	0.770	0.780	0.784
8	0.668	0.695	0.737	0.764	0.778	0.789	0.794	0.791	0.789
9	0.689	0.716	0.748	0.780	0.787	0.796	0.797	0.795	0.793
10	0.715	0.736	0.766	0.791	0.800	0.806	0.804	0.802	0.715
r=4 effects with nonzero means									
6	0.541	0.585	0.624	0.660	0.682	0.693	0.707	0.716	0.720
7	0.591	0.634	0.671	0.692	0.697	0.706	0.719	0.724	0.725
8	0.621	0.664	0.701	0.720	0.726	0.740	0.738	0.734	0.729
9	0.642	0.682	0.715	0.732	0.738	0.743	0.742	0.735	0.731
10	0.660	0.697	0.728	0.744	0.748	0.751	0.746	0.739	0.660
r=5 effects with nonzero means									
6	0.502	0.542	0.583	0.611	0.625	0.635	0.646	0.649	0.645
7	0.547	0.595	0.626	0.638	0.637	0.641	0.649	0.652	0.645

Table 4. Simulation Results: Power of PSE Based Test for Various Values of the Tuning Constants m = the Number of Effects Included in the Initial Scale Estimator and b = the Multiplier Used in Trimming Effects before the Second Scale Estimator

m	b								
	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
8	0.575	0.623	0.656	0.665	0.662	0.668	0.662	0.651	0.642
9	0.586	0.638	0.665	0.672	0.666	0.663	0.661	0.652	0.642
10	0.600	0.647	0.670	0.675	0.667	0.661	0.653	0.645	0.600
r=6 effects with nonzero means									
6	0.456	0.497	0.522	0.539	0.545	0.548	0.548	0.544	0.538
7	0.501	0.546	0.566	0.568	0.552	0.551	0.548	0.542	0.529
8	0.524	0.560	0.581	0.577	0.562	0.559	0.544	0.529	0.516
9	0.529	0.571	0.581	0.567	0.558	0.544	0.531	0.518	0.507
10	0.512	0.541	0.540	0.526	0.517	0.508	0.502	0.498	0.512
r=7 effects with nonzero means									
6	0.413	0.433	0.441	0.446	0.442	0.437	0.431	0.414	0.406
7	0.438	0.467	0.470	0.454	0.435	0.428	0.416	0.403	0.389
8	0.451	0.473	0.471	0.449	0.420	0.402	0.383	0.364	0.353
9	0.419	0.429	0.415	0.392	0.370	0.356	0.353	0.344	0.340
10	0.395	0.384	0.364	0.340	0.327	0.328	0.331	0.333	0.395
r=8 effects with nonzero means									
6	0.342	0.355	0.346	0.336	0.318	0.299	0.281	0.257	0.237
7	0.352	0.364	0.339	0.309	0.280	0.263	0.244	0.225	0.211
8	0.326	0.316	0.285	0.244	0.210	0.190	0.179	0.169	0.166
9	0.282	0.255	0.215	0.183	0.162	0.157	0.158	0.156	0.158
10	0.271	0.235	0.186	0.157	0.146	0.147	0.151	0.156	0.271

Table 5. Simulation Results: Power of the ASE Based Test for Various Values of the Tuning Constants m = the Number of Effects Included in the Initial Scale Estimator and b = the Multiplier Used in Trimming Effects before the Second Scale Estimator

m	b								
	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
r=1 effect with nonzero mean									
6	0.794	0.827	0.868	0.895	0.900	0.918	0.938	0.946	0.953
7	0.840	0.861	0.880	0.910	0.914	0.930	0.942	0.950	0.952
8	0.876	0.882	0.894	0.916	0.924	0.936	0.945	0.952	0.952
9	0.904	0.899	0.906	0.918	0.930	0.940	0.950	0.952	0.956
10	0.921	0.915	0.918	0.925	0.932	0.944	0.953	0.954	0.921
r=2 effects with nonzero means									
6	0.716	0.751	0.796	0.824	0.822	0.839	0.856	0.857	0.847
7	0.753	0.771	0.814	0.843	0.841	0.859	0.862	0.863	0.855
8	0.796	0.798	0.831	0.859	0.858	0.871	0.877	0.872	0.864
9	0.825	0.821	0.838	0.861	0.862	0.876	0.882	0.876	0.866
10	0.846	0.842	0.856	0.871	0.861	0.880	0.883	0.877	0.846
r=3 effects with nonzero means									
6	0.651	0.694	0.739	0.771	0.758	0.764	0.776	0.770	0.756
7	0.694	0.719	0.759	0.786	0.779	0.786	0.787	0.780	0.766
8	0.729	0.746	0.780	0.801	0.790	0.806	0.800	0.789	0.769
9	0.754	0.763	0.785	0.804	0.788	0.813	0.805	0.790	0.768
10	0.781	0.779	0.797	0.811	0.793	0.814	0.810	0.794	0.781
r=4 effects with nonzero means									
6	0.602	0.646	0.684	0.709	0.694	0.690	0.679	0.662	0.638
7	0.645	0.671	0.708	0.725	0.704	0.708	0.699	0.676	0.646
8	0.671	0.696	0.726	0.738	0.716	0.724	0.708	0.684	0.649
9	0.689	0.711	0.731	0.741	0.709	0.722	0.704	0.673	0.641
10	0.699	0.724	0.739	0.729	0.702	0.717	0.690	0.655	0.699
r=5 effects with nonzero means									
6	0.548	0.589	0.624	0.631	0.617	0.596	0.574	0.546	0.514
7	0.585	0.617	0.648	0.647	0.622	0.613	0.588	0.555	0.513

Table 5. Simulation Results: Power of the ASE Based Test for Various Values of the Tuning Constants m = the Number of Effects Included in the Initial Scale Estimator and b = the Multiplier Used in Trimming Effects before the Second Scale Estimator

m	b								
	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
8	0.605	0.635	0.656	0.645	0.618	0.616	0.586	0.546	0.506
9	0.610	0.643	0.652	0.631	0.598	0.590	0.561	0.517	0.480
10	0.608	0.632	0.631	0.589	0.559	0.555	0.512	0.473	0.431
r=6 effects with nonzero means									
6	0.487	0.523	0.540	0.533	0.514	0.493	0.448	0.411	0.376
7	0.514	0.547	0.558	0.542	0.512	0.493	0.448	0.403	0.360
8	0.528	0.548	0.551	0.519	0.487	0.471	0.425	0.384	0.342
9	0.513	0.533	0.515	0.474	0.446	0.420	0.377	0.334	0.298
10	0.430	0.431	0.405	0.354	0.326	0.315	0.282	0.251	0.430
r=7 effects with nonzero means									
6	0.415	0.432	0.432	0.414	0.383	0.364	0.324	0.284	0.251
7	0.423	0.441	0.433	0.397	0.371	0.355	0.306	0.260	0.226
8	0.424	0.421	0.399	0.356	0.324	0.302	0.259	0.216	0.189
9	0.333	0.328	0.294	0.252	0.228	0.208	0.182	0.158	0.145
10	0.246	0.228	0.195	0.158	0.140	0.132	0.127	0.120	0.246
r=8 effects with nonzero means									
6	0.326	0.330	0.309	0.279	0.249	0.236	0.208	0.176	0.149
7	0.313	0.304	0.275	0.240	0.212	0.197	0.160	0.136	0.117
8	0.245	0.227	0.198	0.161	0.142	0.132	0.110	0.097	0.088
9	0.160	0.140	0.117	0.090	0.080	0.076	0.071	0.070	0.072
10	0.112	0.101	0.079	0.064	0.055	0.058	0.061	0.066	0.112

Table 6. Simulation Results: Power of the TSE Based Test for Various Values of the Tuning Constant m = the Number of Effects Included in the Scale Estimate

r = number of effects with nonzero means	m				
	6	7	8	9	10
1	0.855	0.890	0.914	0.923	0.930
2	0.772	0.808	0.829	0.844	0.859
3	0.708	0.740	0.760	0.774	0.784
4	0.648	0.674	0.689	0.699	0.704
5	0.576	0.601	0.613	0.611	0.595
6	0.493	0.511	0.512	0.494	0.428
7	0.399	0.401	0.382	0.326	0.245
8	0.291	0.272	0.220	0.158	0.113

Legends for Figures

Figure1. Difference in power from the overall optimum for tests based on the PSE, and ASE for the case that the nonnull effects have mean value (f) of 4 (recall that there are k=15 effects): tests based on the ASE with tuning constants ....., (m=10,b=2.5) and ---, (m=8,b=2.5); the PSE test with tuning constants _____, (m=8,b=2.5) and _ -_ -_, (m=7,b=1.25); and for the TSE test _ _ _ _, m=9 where m is the number of effects used to calculate the initial scale estimate and b is the multiplier used in trimming large effects prior to calculating the final scale estimate. All differences have standard errors less than 0.01.

Figure 2. Pareto plot (a) and half-normal plot (b) for example B&M-I. The vertical lines are the estimated scales multiplied by the α=0.05 empirical critical values for the following methods: _____, PSE(q=.5, b=2.5); ....., ASE(q=.5, b=2.5); -----, PSE(q=0.467, b=1.25); and _ _ _ _, TSE(q=.6) where q is the proportion of effects included in the initial scale estimator and b is the multiplier used in trimming large effects prior to calculating the final scale estimate. The slope of the line on the half-normal plot is based on the PSE(q=0.5, b=2.5).

Figure 3. Pareto plot (a) and half-normal plot (b) for example B&M-IV. The vertical lines are the estimated scales multiplied by the α=0.05 empirical critical values for the following methods: _____, PSE(q=.5, b=2.5); ....., ASE(q=.5, b=2.5); -----, PSE(q=0.467, b=1.25); and _ _ _ _, TSE(q=.6) where q is the proportion of effects included in the initial scale estimator and b is the multiplier used in trimming large effects prior to calculating the final scale estimate. The slope of the line on the half-normal plot is based on the PSE(q=0.5, b=2.5).

Figure 4. Pareto plot (a) and half-normal plot (b) for the example from Kannan et al (1988). The vertical lines are the estimated scales multiplied by the α=0.05 empirical crit-

ical values for the following methods: _____, PSE(q=.5, b=2.5); ....., ASE(q=.5, b=2.5); ----, PSE(q=0.467, b=1.25); and _ _ _ _, TSE(q=.6) where q is the proportion of effects included in the initial scale estimator and b is the multiplier used in trimming large effects prior to calculating the final scale estimate. The slope of the line on the half-normal plot is based on the PSE(q=0.5, b=2.5).

Figure 5. Active contrast plots for the Kannan et al. (1988) example with priors of (a) α=.1, k=5, (b) α=.2, k=10, and (c) α=.3, k=15 where α is the proportion of effects believed to be "active" and k is a scale factor.

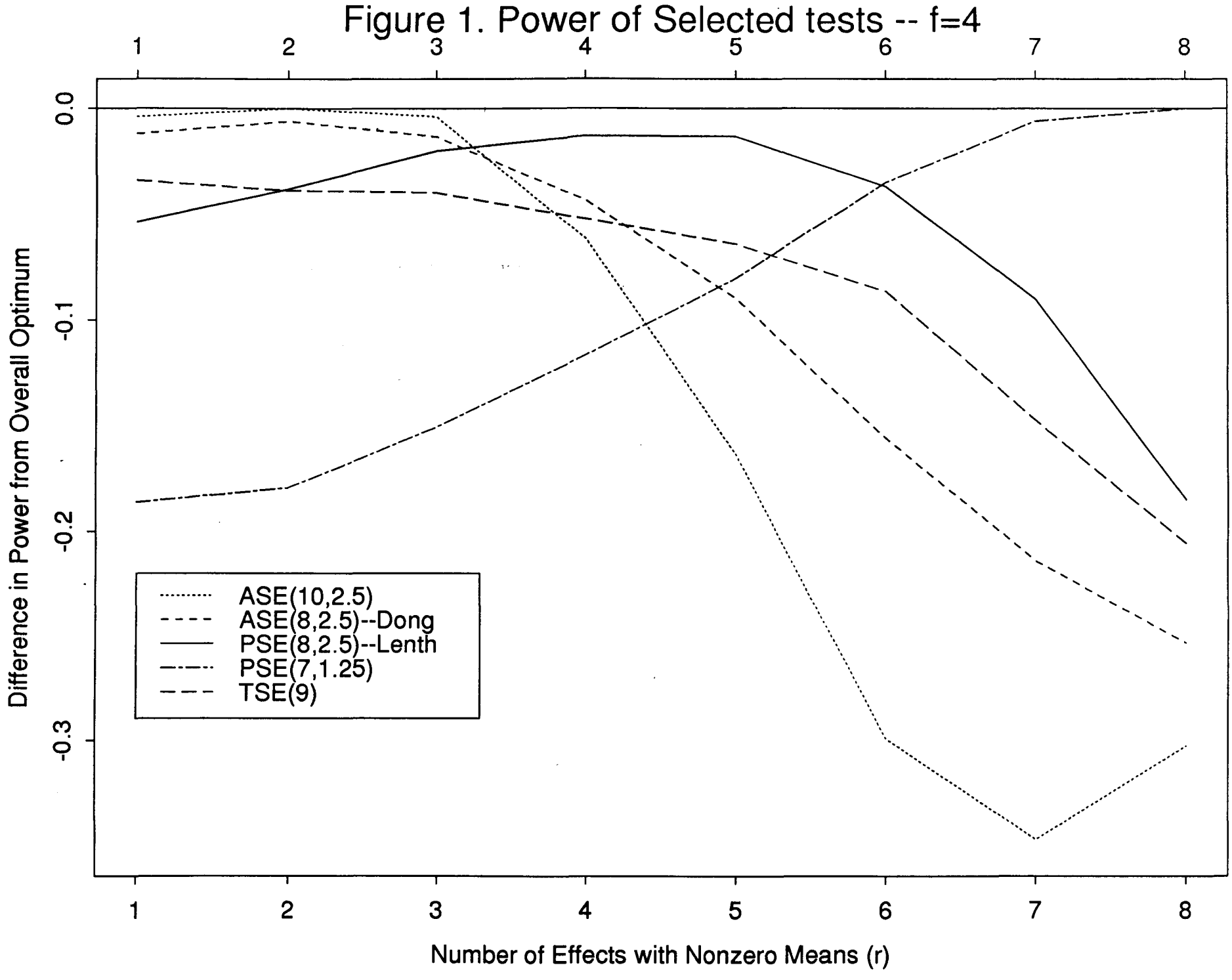
Figure 1. Power of Selected tests -- f=4
1
2
3
4
5
6
7
8
0.0
-0.1
-0.2
-0.3
Difference in Power from Overall Optimum
ASE(10,2.5)
ASE(8,2.5)--Dong
PSE(8,2.5)--Lenth
PSE(7,1.25)
TSE(9)
Number of Effects with Nonzero Means (r)

Figure 2. Graphical Analysis of Example B&M-I

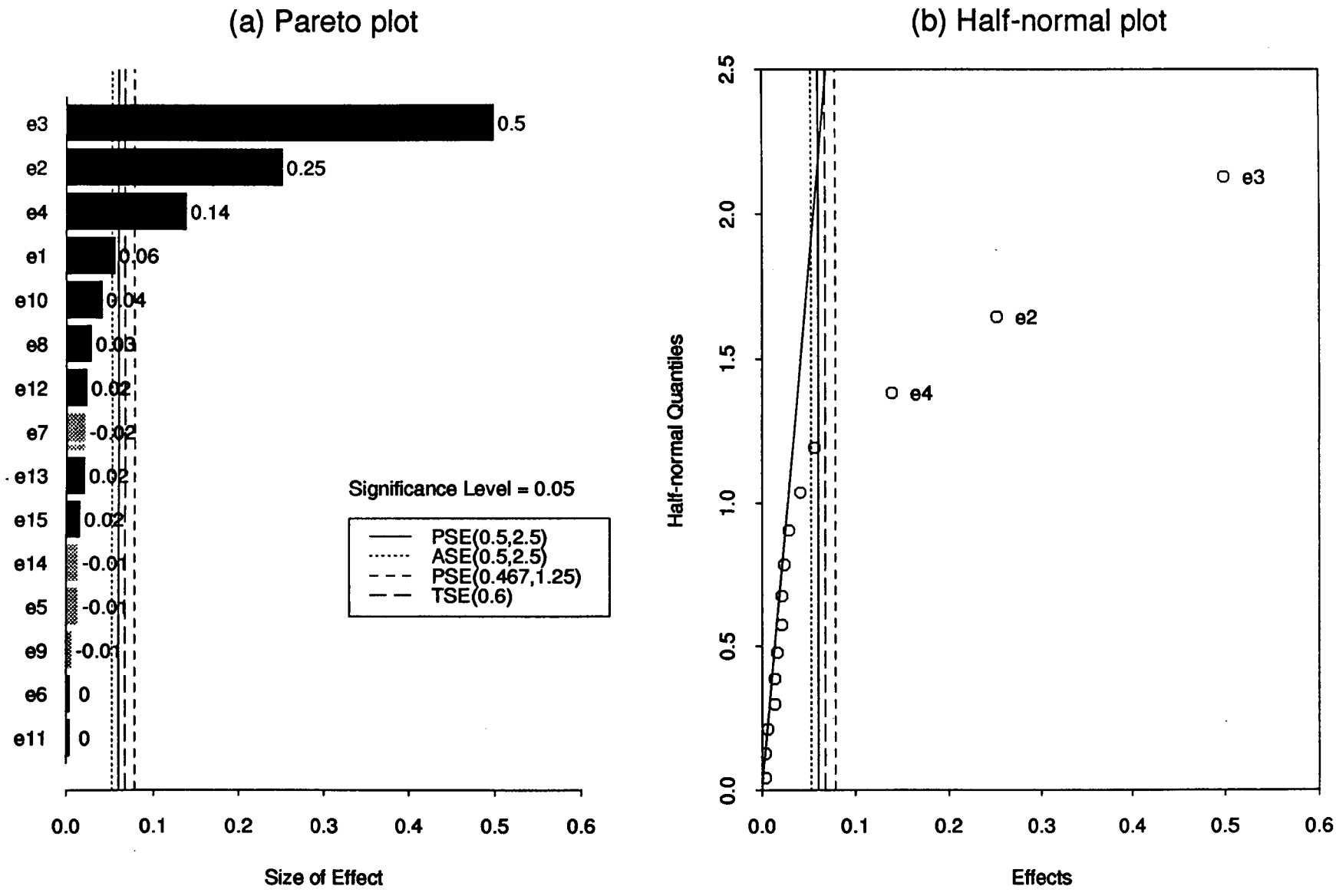

Figure 3. Graphical Analysis of Example B&M-IV

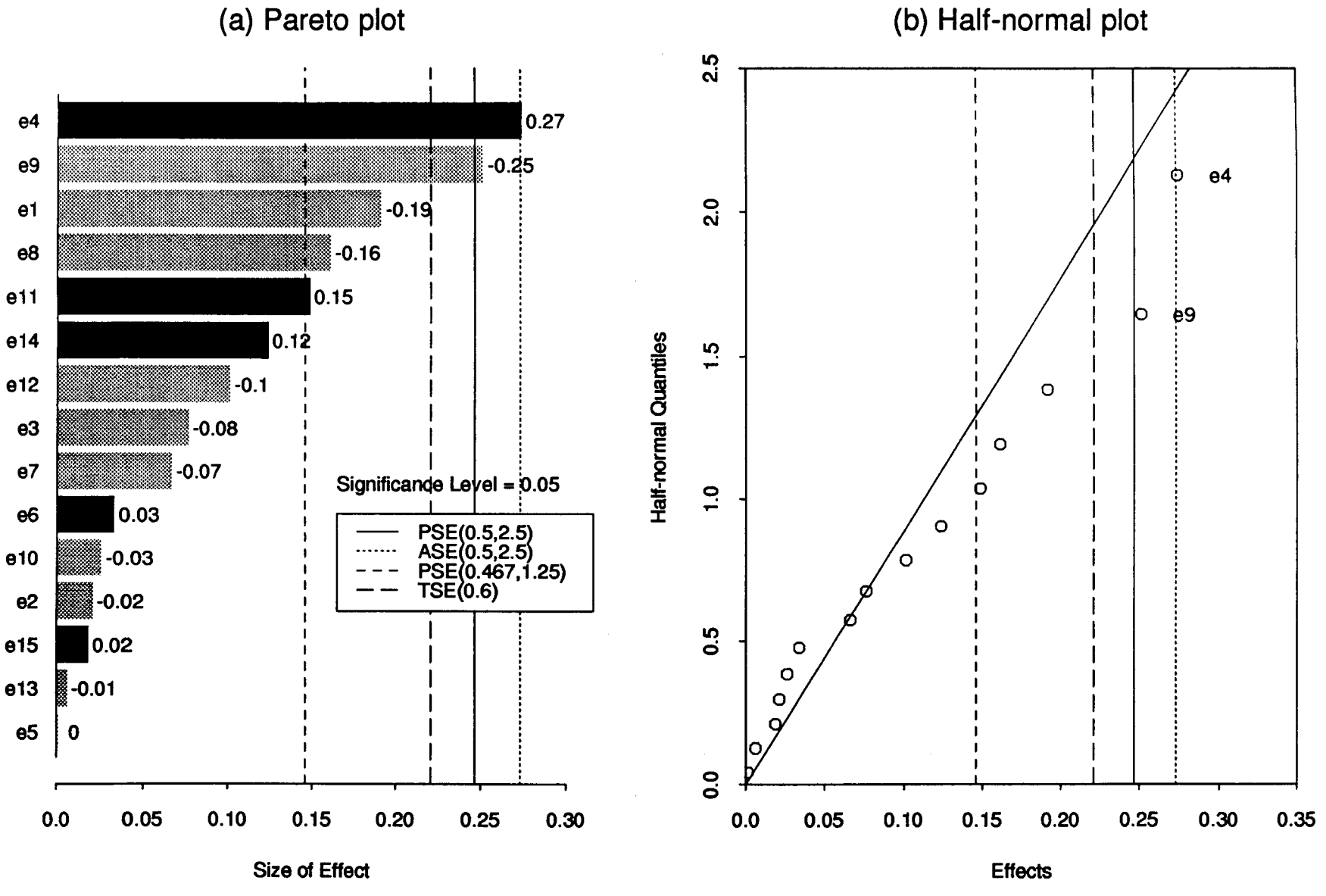

Figure 4. Graphical Analysis of Example from Kannan et al. (1988)

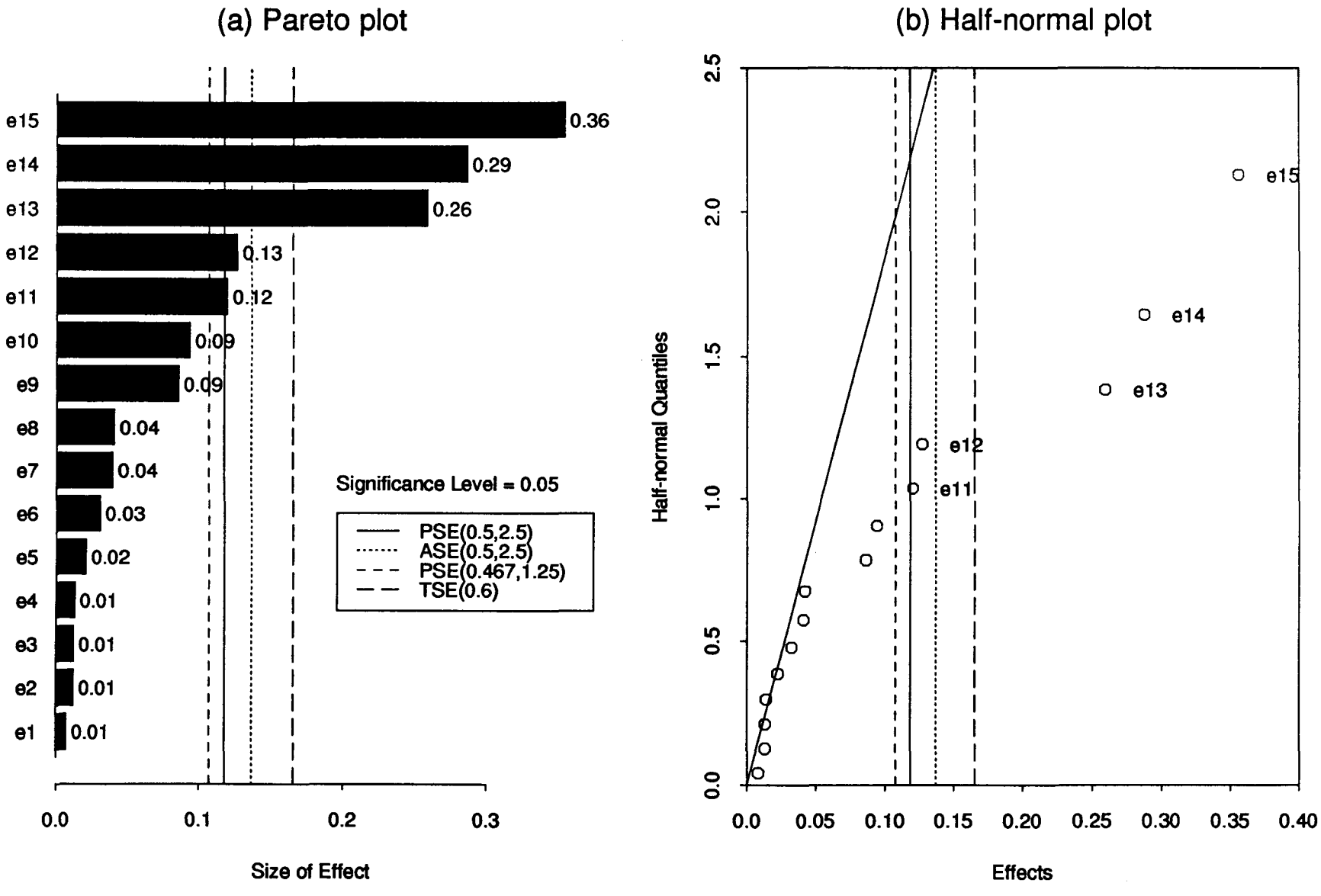

Figure 5. Active Contrast Plots for Kannan et al. (1988)

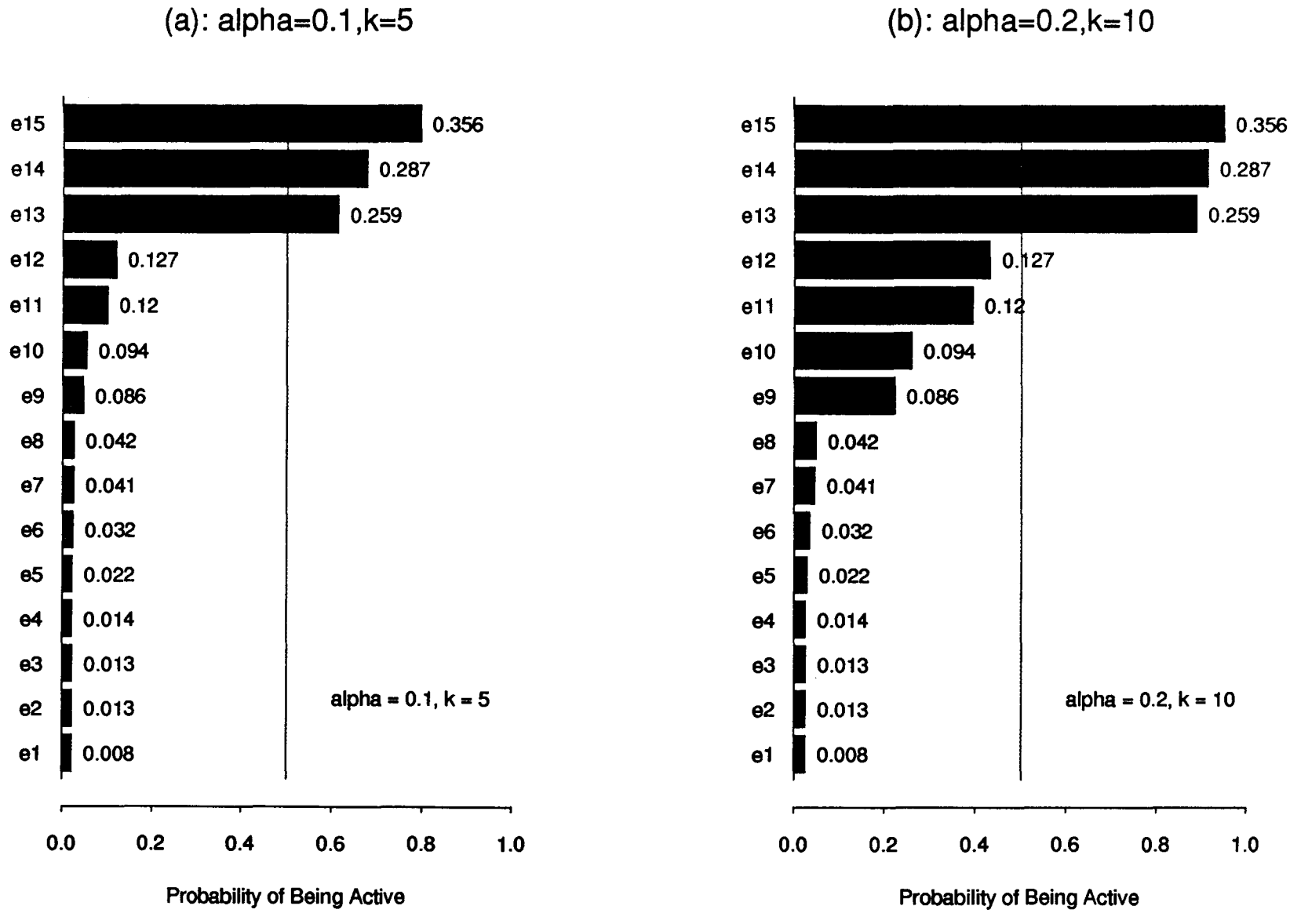

Figure 5. Active Contrast Plots for Kannan et al. (1988)