JIRSS (2023)
Vol. 22, No. 01, pp 161-174
DOI: 10.22034/jirss.2023.2006195.1029

A Note on the Non-Existence of a Best Test for Analyzing Unreplicated Two-Level Factorial Designs

Soltan Mohammad Sadooghi-Alvandi[1], **Mahmood Kharrati-Kopaei** [1]

[1] Department of Statistics, Shiraz University, Shiraz, Iran.

Received: 03/07/2023, Accepted: 23/12/2023, Published online: 15/06/2024

Abstract. The usual t-test or F-test can not be used to analyze unreplicated two-level factorial designs, since all the observations are used to estimate the factor effects and no observation is left to estimate the error variance. To overcome this difficulty, various procedures have been proposed in the literature and several simulation studies have been carried out to compare the performance of these methods. The results of these studies have been inconclusive, and no test is widely accepted as a "best" test. In this paper, we present results that show theoretically that no test has high power against all possible alternatives; i.e. no test can detect all patterns of active effects. Therefore, in the absence of any prior information concerning active and inactive effects, no test can be preferred to any other test based on power and the choice of a test should be based on other considerations, such as ease of use, control of individual or experimental error rate, the purpose of the experiment, etc.

Keywords. Active Effect, EER, Error Rate, IER, Inactive Effect, Power.

MSC: 62F99, 62K15.

1 Introduction

The standard orthogonal model for the analysis of an unreplicated 2^k factorial design may be written a

$$\boldsymbol{y} = \mu \boldsymbol{x}_0 + \sum_{j=1}^{m} \beta_j \boldsymbol{x}_j + \boldsymbol{\epsilon}, \tag{1.1}$$

Soltan Mohammad Sadooghi-Alvandi (smsa51@hotmail.com)
Corresponding Author: Mahmood Kharrati-Kopaei (mkharati@shirazu.ac.ir)

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of observations ($n = 2^k$), and μ is the general mean, and $\beta_1, \ldots, \beta_m$ are the factorial effects ($m = n - 1$), and $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are known orthogonal vectors with elements -1 and $+1$ ($\boldsymbol{x}_0$ is the n-dimensional vector of ones), and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is a vector of random errors, assumed to be independent normal variables with mean zero and variance σ^2 (hereafter $N(0, \sigma^2)$). The main parameters of interest are $\beta_1, \ldots, \beta_m$, and μ and σ^2 are nuisance parameters. Note that the error terms in the model (1.1) have a common variance σ^2. Therefore, we do not consider the heterogeneous model here.

Since model (1.1) is saturated, all of the observations are used to estimate $\mu, \beta_1, \ldots, \beta_m$ and no observation is left to estimate σ^2. Therefore, the usual t-test or F-test cannot be used to test if the factor effects, β_i, are zero or not. Various methods have been proposed to overcome this problem. The first method, introduced by Daniel (1959), is to assess the significance of the estimated effects, $\hat{\beta}_i$, using normal or half-normal plots: Estimated effects that are approximately near the straight line on the plot are considered inactive effects (the corresponding β_i's can be assumed zero) and those which appear at a distance from the line are considered as active effects (the corresponding β_i's are non-zero). The main disadvantage of this graphical method is that the interpretation of the plot is subjective. This has motivated attempts to propose various objective methods for detecting active and inactive effects: see, e.g., (Birnbaum, 1959, 1961; Holms and Berrettoni, 1969; Zahn, 1975a,b; Seheult and Tukey, 1982; Box and Meyer, 1986; Johnson and Tukey, 1987; Voss, 1988; Benski, 1989; Bissell, 1989, 1992; Lenth, 1989; Berk and Picard, 1991; Juan and Pena, 1992; Le and Zamar, 1992; Loh, 1992; Dong, 1993; Schneider et al., 1993; Benski and Cabau, 1995; Haaland and O'Connell, 1995; Venter and Steel, 1996; Loughin and Noble, 1997; Hamada and Balakrishnan, 1998; Voss, 1999; Al-Shiha and Yang, 1999, 2000; Ye et al., 2001; Aboukalam and Al-Shiha, 2001; Chen, 2003; Chen and Kunert, 2004; Miller, 2005; Voss and Wang, 2006; Wu and Wang, 2007; Kharrati-Kopaei and Shenavari, 2023). For excellent reviews of the different methods, see (Hamada and Balakrishnan, 1998) and (Chen, 2003).

Several simulation studies have been carried out to compare the performance of these methods. An extensive evaluation of the most popular methods can be found in Hamada and Balakrishnan (1998) and Chen (2003); see also (Haaland and O'Connell, 1995; Benski and Cabau, 1995; Al-Shiha and Yang, 1999), and (Chen and Kunert, 2004). The results of these simulation studies have been inconclusive, and the situation is best summarized by Haaland, in his comment in the discussion part of Hamada and Balakrishnan's paper (1998, Comments, pp. 32, 33, 35): *"First, in my opinion, a purely statistical (that is, objective) identification of the active effects is not possible. The "vital few" and "trivial many" will in general be obvious regardless of the test method used. The "statistical in-between" effects will always be a problem (as the authors note) due to a combination of lack of power and the absence of an omnibus test that works over a wide range of the number of active effects. ... There are many competing test methods, which are difficult to distinguish among based on performance, and it is easy to come to wrong conclusions if a careful comparison is not made. The choice of a test is further complicated by the fact that there is no one test that*

performs well over a wide range of numbers of active effects."

We note that the claims such as *"there is no test that performs well over a wide range of number of active effects"* are based on simulation studies. In addition, it is routinely accepted (as far as we know without any theoretical justification) that the existing tests are only useful under the assumption of effect sparsity. It seems to us that these claims must be true intuitively since the number of unknown parameters exceeds the number of observations in the model (1.1). However, the main contribution of this paper is that we prove those claims theoretically with simple math. That is, in Section 3, we present theoretical results that prove that without the assumption of effect sparsity, no test has power higher than the level of the test.

The organization of the paper is as follows. In Section 2, we briefly review some points that are important when assessing and comparing different methods of detecting active and inactive effects. In Section 3, we present our main non-existence result. In Section 4, we illustrate our theoretical results numerically. In the final section, we make some recommendations for applying a test procedure based on considerations such as ease of use, control of individual or experimental error rate, the purpose of the experiment, etc.

2 Error Rate and Power

For clarity, in this section, we briefly review some points that are important when assessing and comparing different methods of detecting active and inactive effects. Note that we are dealing with a multiple-testing problem. The first point to consider is the error rate; i.e., how the size of the test is controlled. Two kinds of error rates are usually used. The first is the individual error rate (IER); which is the expected fraction of inactive effects that are declared active. The second is the experimental error rate (EER); which is the probability that at least one inactive effect is declared active. Each error rate, however, can be used in a "weak" or "strong" sense. For clarity, let

$$H_i : \beta_i = 0 \quad \text{and} \quad H_i' : \beta_i \neq 0, \quad i = 1, \ldots, m, \tag{2.1}$$

and let $H_0 = \cap_{i=1}^{m} H_i$ denote the global hypothesis of no active effect:

$$H_0 : \beta_i = 0, \quad \text{for all} \quad i = 1, \ldots, m. \tag{2.2}$$

(1) Weak individual error rate (WIER) control. We say that IER is controlled at level α in a weak sense if

$$\frac{1}{m} \sum_{i=1}^{m} \Pr\{\text{Reject } H_i \mid \beta_1 = 0, \ldots, \beta_m = 0\} \leq \alpha, \tag{2.3}$$

see, e.g., (Hamada and Balakrishnan, 1998; Ye et al., 2001), and (Chen, 2003). Note that if $\Pr\{\text{Reject } H_i \mid \beta_1 = 0, \ldots, \beta_m = 0\} \leq \alpha$ for all $i = 1, \ldots, m$, then WIER $\leq \alpha$.

(2) Strong individual error rate (SIER) control. Let J be a subset of $\{1, \ldots, m\}$ and $|J|$ denote the cardinality of J. We say that IER is controlled at level α in a strong sense if

$$\frac{1}{|J|} \sum_{i \in J} \Pr\{\text{Reject } H_i \mid \beta_j = 0, \text{ for all } j \in J\} \leq \alpha, \text{for all } J \subset \{1, \ldots, m\}, \tag{2.4}$$

see (Voss, 1999). Note that (2.3) is a special case of (2.4) with $J = \{1, \ldots, m\}$.

(3) Weak experimental error rate (WEER) control. We say that the EER is controlled at level α in a weak sense if

$$\Pr\{\text{Reject at least one } H_i \mid \beta_1 = 0, \ldots, \beta_m = 0\} \leq \alpha, \tag{2.5}$$

see, e.g., (Ye et al., 2001), and (Westfall and Young, 1993, pp. 9-10), and (Hamada and Balakrishnan, 1998).

(4) Strong experimental error rate (SEER) control. We say that the EER is controlled at level α in a strong sense if

$$\Pr\{\text{Reject } H_i \text{ for some } j \in J \mid \beta_j = 0, \text{ for all } j \in J\} \leq \alpha, \text{ for all } J \subset \{1, \ldots, m\},$$

see, e.g., (Voss, 1999) and (Westfall and Young, 1993, pp. 9-10). Note that, in this case, (2.5) also holds.

As for power, two kinds of power, referred to as Power I and Power II, are usually used; for a discussion of other types of power, see (Chen, 2003, pp. 67, 68).

(I) Power I is the probability of rejecting the global hypothesis H_0 when at least one effect is active

$$\text{Power I} = \Pr\{\text{Reject at least one } H_i \mid \beta_j \neq 0, \text{ for some } j = 1, \ldots, m\};$$

see, e.g., (Hamada and Balakrishnan, 1998) and (Ye et al., 2001). Note that the i for which H_i is rejected need not be the same as the j for which $\beta_j \neq 0$.

(II) Power II is the expected fraction of active effects that are declared active; see, e.g., (Hamada and Balakrishnan, 1998; Ye et al., 2001; Chen and Kunert, 2004). Let J denote the set of active effects: $J = \{j : \beta_j \neq 0\}$. Then Power II may be written as

$$\text{Power II} = \frac{1}{|J|} \sum_{i \in J} \Pr\{\text{Reject } H_i \mid \beta_j \neq 0, \text{ for } j \in J\},$$

where $|J|$ denotes the cardinality of J.

3 Main Rresult: Non-Existence of a Best Test

In this section, we first show that for any test procedure that controls the error rate at level α its "average power" is also at most α when the effects are generated from $N(0, \sigma_\beta^2)$. By 'average power', we refer to generating the values of $\beta_1, \ldots, \beta_m$ from the normal distribution $N(0, \sigma_\beta^2)$ and then calculating as well the expectation of Power I and Power II, over the distribution of $\beta_1, \ldots, \beta_m$. The same approach was used, in another context, in Kharrati-Koapei and Sadooghi-Alvandi (2007). Then, using this result, we show that no test has a power higher than α without the assumption of effect sparsity. Note that consideration of "average power", when the effects are generated from a normal distribution, is used merely as a "tool" for proving our main non-existence result.

Before presenting our results, we need some preliminaries. It will be assumed that any test procedure for detecting active effects is based on $\hat{\beta}_i$'s (the usual least squares estimates of β_i's). This is indeed the case for all procedures proposed in the literature, and may be theoretically justified (using the Conditionality Principle or Invariance Principle) as follows: It is easily verified that, for the saturated and orthogonal model (1.1), the least square (maximum likelihood) estimators of $\mu, \beta_1, \ldots, \beta_m$ are given by

$$\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\beta}_i = \frac{1}{n}\boldsymbol{x}_i'\boldsymbol{y}, \quad i = 1, \ldots, m,$$

and that, $\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_m$ are independent normal variables, with $\hat{\mu} \sim N(\mu, \sigma^2/n)$ and $\hat{\beta}_i \sim N(\beta_i, \sigma^2/n)$. It can also be shown that there is a one-to-one relationship between $(y_1, \ldots, y_n)$ and $(\hat{\mu}, \hat{\beta}_1, \ldots, \hat{\beta}_m)$. We therefore "assume" that inferences about $\beta_1, \ldots, \beta_m$ are to be based only on $\hat{\beta}_1, \ldots, \hat{\beta}_m$. This "assumption" can be justified by the Conditionality Principle, since the distribution of $\hat{\mu}$ does not depend on $\beta_1, \ldots, \beta_m$, and the conditional distribution of the observations, given $\hat{\mu}$, depends on $y_1, \ldots, y_n$ only through $\hat{\beta}_1, \ldots, \hat{\beta}_m$. It can also be justified by the Invariance Principle, since the problem is invariant under translations (shifts in location) and $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are invariant under translations. More specifically, if the observations are transformed to $y_i^* = y_i + c$ (where c is an arbitrary real number), then the model for $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$ is $\boldsymbol{y}^* = \mu^*\boldsymbol{x}_0 + \sum_{j=1}^m \beta_j \boldsymbol{x}_j + \boldsymbol{\epsilon}$, where $\mu^* = \mu + c$, and it is easily verified that $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are maximal invariant under this group of transformations. In what follows, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ and let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ denote the least squares estimator of $\boldsymbol{\beta}$. When considering "average power", it is assumed that $\beta_1, \ldots, \beta_m$ are independent $N(0, \sigma_\beta^2)$ variables; i.e. $\boldsymbol{\beta} \sim N_m(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_m)$.

Lemma 3.1. *Suppose that $g(\hat{\boldsymbol{\beta}})$ is a function of $\hat{\boldsymbol{\beta}}$ and let $\phi(\boldsymbol{\beta}, \sigma^2) = E\left(g(\hat{\boldsymbol{\beta}}) \mid \boldsymbol{\beta}; \sigma^2\right)$; i.e., more specifically, $\phi(\boldsymbol{\beta}_0, \sigma_0^2) = E\left(g(\hat{\boldsymbol{\beta}}) \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0; \sigma^2 = \sigma_0^2\right)$. If $\boldsymbol{\beta} \sim N_m(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_m)$, then*

$$E_{\boldsymbol{\beta}}(\phi(\boldsymbol{\beta}, \sigma^2) = \phi(\boldsymbol{0}; \sigma^2 + n\sigma_\beta^2).$$

(That is, the expectation of $\phi(\boldsymbol{\beta}, \sigma^2)$, over $\boldsymbol{\beta}$, is the same as the expectation of $g(\hat{\boldsymbol{\beta}})$ when $\boldsymbol{\beta} = \boldsymbol{0}$ and the error variance is $\sigma^2 + n\sigma_\beta^2$.)

Proof. As noted before, it is easily verified that the least squares estimators of β_i's, under model (1.1), are independent and normally distributed with mean β_i and variance σ^2/n; i.e. the (conditional) distribution of $\hat{\boldsymbol{\beta}}$ (given $\boldsymbol{\beta}$) is $N_m(\boldsymbol{\beta}, (\sigma^2/n)\boldsymbol{I}_m)$. Thus

$$\phi(\boldsymbol{\beta}, \sigma^2) = E_{\boldsymbol{X}}\left(g(\boldsymbol{X})\right), \text{where } \boldsymbol{X} \sim N_m(\boldsymbol{\beta}, (\sigma^2/n)\boldsymbol{I}_m). \tag{3.1}$$

Now, it can be shown that if $\boldsymbol{\beta} \sim N_m(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_m)$ and $\hat{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \sim N_m(\boldsymbol{\beta}, (\sigma^2/n)\boldsymbol{I}_m)$, then the unconditional distribution of $\hat{\boldsymbol{\beta}}$ is $N_m(\boldsymbol{0}, \sigma_\beta^2 + (\sigma^2/n)\boldsymbol{I}_m)$. Note that $\phi(\boldsymbol{\beta}, \sigma^2)$ is the conditional expectation of $g(\hat{\boldsymbol{\beta}})$ over the distribution of $\hat{\boldsymbol{\beta}} \mid \boldsymbol{\beta}$. Therefore, by using the properties of conditional expectation, we have

$$\begin{aligned} E_{\boldsymbol{\beta}}\left(\phi(\boldsymbol{\beta}, \sigma^2)\right) &= E_{\boldsymbol{\beta}}\left(E_{\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}}\left(g\left(\hat{\boldsymbol{\beta}}\right) \mid \boldsymbol{\beta}, \sigma^2\right)\right) \\ &= E_{\hat{\boldsymbol{\beta}}}\left(g(\hat{\boldsymbol{\beta}})\right) = \phi(\boldsymbol{0}; \sigma^2 + n\sigma_\beta^2), \end{aligned}$$

where the last equality follows from (3.1), and the fact that the unconditional distribution of $\hat{\boldsymbol{\beta}}$ is $N_m(\boldsymbol{0}, \sigma_\beta^2 + (\sigma^2/n)\boldsymbol{I}_m)$. This completes the proof. □

Now, we can show that for any test procedure that controls Type one error at level α, the value of its "average power" is at most α when all effects are active. Recall the assumption, justified earlier, that any test procedure for testing β_i's in model (1.1) is to be based on $\hat{\beta}_i$'s. In what follows, $T(\hat{\boldsymbol{\beta}})$ is a test statistic and C_i denotes the critical region for testing $H_i : \beta_i = 0$ (i.e. reject H_i if $T(\hat{\boldsymbol{\beta}}) \in C_i$) and C denotes the critical region for testing the global hypothesis $H_0 = \cap_{i=1}^m H_i$ (i.e. reject H_0 if $T(\hat{\boldsymbol{\beta}}) \in C$). Also, $\Pi_I(\boldsymbol{\beta}, \sigma^2)$ and $\Pi_{II}(\boldsymbol{\beta}, \sigma^2)$ denote Power I and Power II of the test, respectively.

Proposition 3.1 (Power I)**.** *Suppose that the test procedure $T(\hat{\boldsymbol{\beta}})$ controls EER at level α (at least in the weak sense); i.e.,*

$$Pr\{T(\hat{\boldsymbol{\beta}}) \in C \mid \boldsymbol{\beta} = \boldsymbol{0}; \sigma^2\} \leq \alpha, \text{ for all } \sigma^2. \tag{3.2}$$

If $\boldsymbol{\beta} \sim N_m(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_m)$, then $E_{\boldsymbol{\beta}}\left(\Pi_I(\boldsymbol{\beta}, \sigma^2)\right) \leq \alpha$. That is, the average power of the test is at most α.

Proof. Let $g(\hat{\boldsymbol{\beta}}) = I_{\{T(\hat{\boldsymbol{\beta}}) \in C\}}$, where I_A denote the indicator function of event A; i.e.

$$g(\hat{\boldsymbol{\beta}}) = \begin{cases} 1 & \text{if } T(\hat{\boldsymbol{\beta}}) \in C \\ 0 & \text{if } T(\hat{\boldsymbol{\beta}}) \notin C. \end{cases}$$

Then $\Pi_I(\boldsymbol{\beta}, \sigma^2) = \Pr\{T(\hat{\boldsymbol{\beta}}) \in C \mid \boldsymbol{\beta}; \sigma^2\} = E\left(g(\hat{\boldsymbol{\beta}}) \mid \boldsymbol{\beta}; \sigma^2\right)$. Using Lemma 3.1, we have

$$\begin{aligned} E_{\boldsymbol{\beta}}\left(\Pi_I(\boldsymbol{\beta}, \sigma^2)\right) &= \Pi_I(\boldsymbol{0}; \sigma^2 + n\sigma_\beta^2) \\ &= \Pr\{T(\hat{\boldsymbol{\beta}}) \in C \mid \boldsymbol{\beta} = \boldsymbol{0}; \sigma^2 + n\sigma_\beta^2\} \leq \alpha, \end{aligned}$$

where the last inequality follows from (3.2).

□

Proposition 3.2 (Power II)**.** *Suppose that the test procedure $T(\hat{\boldsymbol{\beta}})$ controls IER at level α (at least in the weak sense); i.e.,*

$$\frac{1}{m}\sum_{i=1}^{m} Pr\{T(\hat{\boldsymbol{\beta}}) \in C_i \mid \boldsymbol{\beta} = \mathbf{0}; \sigma^2\} \leq \alpha, \text{ for all } \sigma^2, \tag{3.3}$$

If $\boldsymbol{\beta} \sim N_m(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}_m)$, then $E_{\boldsymbol{\beta}}\left(\Pi_{II}(\boldsymbol{\beta}, \sigma^2)\right) \leq \alpha$. That is, the average power of the test is at most α.

Proof. Let $g(\hat{\boldsymbol{\beta}})$ denote the fraction of active effects that are declared active; i.e., $g(\hat{\boldsymbol{\beta}}) = \frac{1}{m}\sum_{i=1}^{m} I_{\{T(\hat{\boldsymbol{\beta}}) \in C_i\}}$ (Note that when $\beta_1, \ldots, \beta_m$ are $N(0, \sigma_\beta^2)$, then all the effects are active with probability 1). Then $\Pi_{II}(\boldsymbol{\beta}, \sigma^2) = E\left(g(\hat{\boldsymbol{\beta}}) \mid \boldsymbol{\beta}; \sigma^2\right)$. Using Lemma 3.1, we have

$$\begin{aligned} E_{\boldsymbol{\beta}}\left(\Pi_{II}(\boldsymbol{\beta}, \sigma^2)\right) &= \Pi_{II}(\mathbf{0}; \sigma^2 + n\sigma_\beta^2) \\ &= \frac{1}{m}\sum_{i=1}^{m} \Pr\{T(\hat{\boldsymbol{\beta}}) \in C_i \mid \boldsymbol{\beta} = \mathbf{0}; \sigma^2 + n\sigma_\beta^2\} \leq \alpha, \end{aligned}$$

where the last inequality follows from (3.3).

□

Note 1. Note that if equality is achieved in either (3.2) or (3.3), then the 'average power' will also be equal to α.

Now, we can present our main non-existing result.

Theorem 3.1 (Non-existence of a best test)**.** *There is no test that controls EER (IER) at level α and has power I (power II) higher than α without assuming the sparsity assumption.*

Proof. (Proof is by contradiction). Suppose that there is such a test; i.e., there is a test that has power I (power II) higher than α without assuming the sparsity assumption. Since we do not consider the assumption of effect sparsity, the effects can be taken from $N(0, \sigma_\beta^2)$ (note that when β's are generated from the normal distribution, all β's are non-zero with probability one). But in this case, the "average power" of the test would exceed α, which contradicts the results of Propositions 3.1 and 3.2. Since the value of α is small, it follows that no test can have high power against all alternatives; i.e., no test can detect all patterns of active effects without the assumption of effect sparsity and this completes the proof. □

It follows from Theorem 3.1 that, in the absence of any prior information concerning active and inactive effects, no test can be preferred to any other test on the basis of power. So the choice of a test should be based on other considerations. We will discuss some of these considerations in Section 5.

4 Numerical Illustration

In this section, we illustrate our theoretical results presented in Section 3 via a simulation study. In our numerical illustration, we considered four recent methods as follows:

CK04. Chen and Kunert (2004) proposed a test statistic, $MaxU_r$, designed for the analysis of the unreplicated factorial designs. Their method has the potential to identify up to $m-1$ active contrasts. According to Chen and Kunert (2004), their method has the advantage that it performs reasonably well for all numbers of active factors of up to 8 in 16-run experiments. We shall refer to this procedure by CK04. This method controls WEER.

LH89. Haaland and O'Connell (1995) have shown that the method proposed by Lenth (1989) provides the superior overall test performance among several robust methods. Also as pointed out by Montgomery (2001, p. 254) it is easy to implement. We used the revised critical values, coefficients, and consistency constants for this test statistic, as suggested by Haaland and O'Connell (1995). We shall refer to this procedure by LH89. This method controls WIER.

VW06. Voss and Wang (2006) proposed an adaptive step-down test procedure that includes as special cases the tests of Berk and Picard (1991) and Voss (1988). They indicated that their test works effectively under effect sparsity, and also noted that the step-down tests have a power advantage over simultaneous confidence intervals and single-step tests. We shall refer to this procedure by VW06. This procedure controls SEER.

WW07. Wu and Wang (2007) proposed a step-up simultaneous test under the assumption of effect sparsity. They supposed that the researcher believes that the number of inactive effects is at least ν, but he/she does not know which effects are inactive. They proposed two step-up procedures, with fixed scaling and with sequential scaling. Based on their simulation results (with $\nu = 7$), they recommended the procedure with sequential scaling. We shall refer to this procedure as WW07. This method controls EER in a strong sense.

In our simulations, we considered a 2^4 factorial design ($n = 16$). For VW06, we used the set $J = \{8, 12\}$, as in their simulation. For WW07, we also used $\nu = 7$. For LH89, we used the tuning constants $q = 0.5$ and $b = 2.5$. We used the following procedure to obtain the "average power" of each test:

Step I. We first generated the values of $\beta_1, \ldots, \beta_m$ at random from the normal distribution $N(0, \sigma_\beta^2)$. (We shall refer to this step as a "case".)

Step II. We then generated the values of $\epsilon_1, \ldots, \epsilon_n$ from the normal distribution $N(0, \sigma^2)$, calculated the values of the observations $y_1, \ldots, y_n$, and applied each test. (Since the value μ does not affect the results, its value was set to zero). For each test, we recorded whether the global hypothesis was rejected or not (for Power I), and the fraction of active effects which were declared active (for Power II).

For each case, Step II was repeated $N = 10,000$ times, and the power was calculated for each test procedure. To calculate the "average power" of the four procedures, the two steps were repeated $M = 1000$ times and the average power, over the 1000 cases, was calculated. The results are shown in Table 1. In this table, the "actual error rate" was calculated based on 100,000 samples when all β's were zero. (Note that only for calculating the "average power", the β's were generated at random.) The main point to note is that, in terms of "average power", the values of Power I and Power II are very close to the actual EER and IER, respectively (as formally shown in Section 3). Also, as expected, EER is much higher than IER and Power II is much lower than Power I. Finally, note that LH89 method, which controls IER= 0.05, has a high EER= 0.39797. This explains the relatively high value of Power I for LH89 (but note that the value of Power II is still close to the actual IER).

Table 1: Estimated "average power" of four methods.

	Overall power		Actual error rate (Based on 100,000 samples)		
Methods	Power I	Power II	EER	IER	Nominal error rate
CK04	0.05041	0.00360	0.04972	0.00355	WEER=0.05
LH98	0.38599	0.05043	0.39797	0.05035	WIER=0.05
WV06	0.05235	0.00960	0.05064	0.00996	SEER=0.05
WW07	0.05167	0.01222	0.04895	0.01371	SEER=0.05

For the graphical presentation of power methods, Steps I and II were repeated to obtain $M = 100$ cases. Figures 1-4 show the results for a 2^4 factorial design ($n = 16$, $m = 15$), with EER= 0.05, IER= 0.05, $\sigma^2 = 1$, and $\sigma^2_\beta = 5$. These results are typical: (i) for each method, the power fluctuated from case to case, with some cases of high power and many cases of low power, (ii) the cases of high power were not the same for the four methods; i.e. a test may perform poorly when another test performed well, and (iii) no method performed better than the other three methods in all cases.

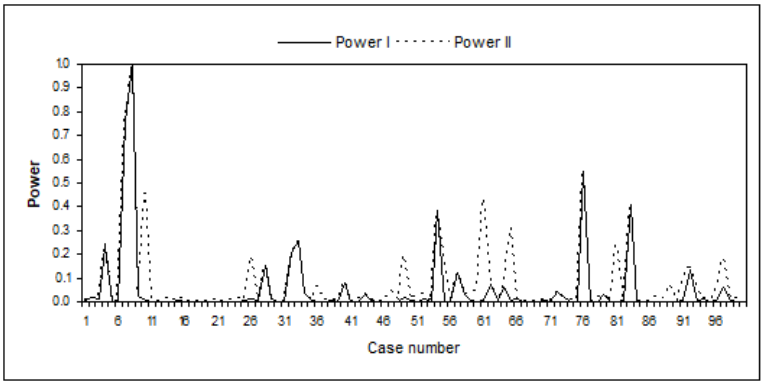

Figure 1: Power I and Power II for Method CK04 with WEER=0.05.

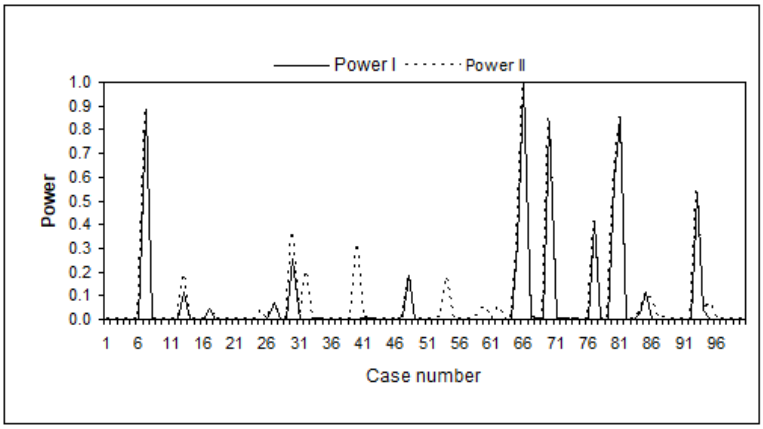

Figure 2: Power I and Power II for Method VW06 with SEER=0.05.

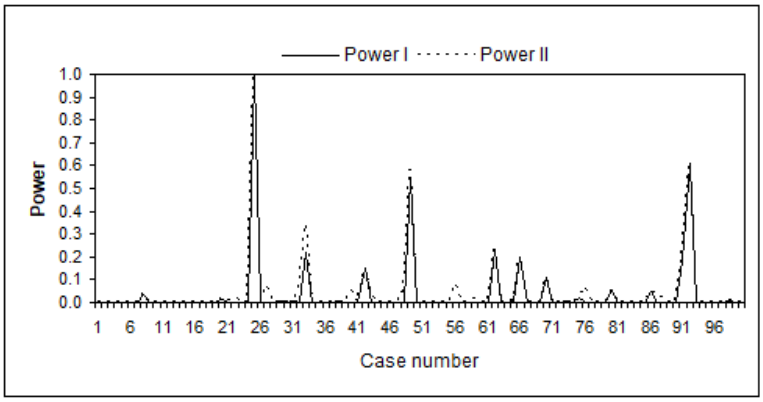

Figure 3: Power I and Power II for Method WW07 with SEER=0.05.

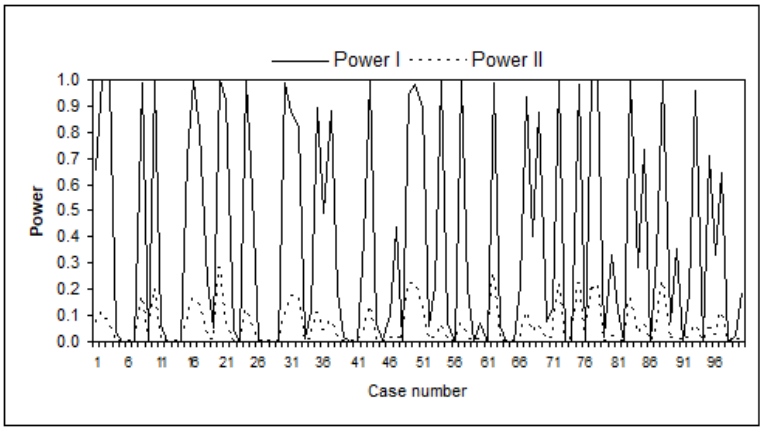

Figure 4: Power I and Power II for Method LH89, with WIER=0.05.

5 Remarks

As noted in the introduction, many procedures have been proposed for the analysis of unreplicated two-level factorial designs, but no test is widely accepted as a "best" test. Our results show that there can be no test that performs well for all configurations of active and inactive effects, and that, in the absence of any prior information concerning active and inactive effects (including sparsity information), no test can be preferred to any other test based on power (thus giving theoretical support to the comments by Haaland, quoted in Section 1). So the choice of a test should be based on other considerations such as simplicity, ease of use, availability (and accuracy) of critical values, and prior information about the pattern of active effects.

Hamada and Balakrishnan (1998, pp. 21-23) compared several methods mostly based on the number of active effects, and made some recommendations as follows: For up to six active effects, the methods proposed by (Daniel, 1959; Schneider et al., 1993; Zahn, 1975a; Berk and Picard, 1991), and half-normal version of Loh (1992) perform well, although other methods are competitive if one has a good idea about how many active effects there are. For example, for eight active effects, the methods of Daniel (1959) and Zahn (1975a) would not be expected to do well since they assume that there would be no more than six active effects. The power of Bissell's (1992) method seriously degrades when there are many active effects so this method is not recommended. For the same reasons, the method of Benski (1989) is also not recommended. According to Chen and Kunert (2004), their method has the advantage that it performs reasonably well for all numbers of active factors of up to 8 in 16-run experiments.

As noted by Montgomery (2001, p. 254), Lenth's method may be preferred to other methods since it is easy to implement. However, as noted in Section 3, while this method controls IER, it can have a large EER. Note that unreplicated designs are often used in the screening stage, to identify the main factors for further studies, in which case *"misjudging inactive effects as being active may not be a serious mistake . . . because we have the opportunity of making a confirmation experiment to check our predictions"* Chen (2003, p.68). Therefore, in such cases, it seems that methods that control IER are more appropriate (even though they may have large EER).

If control of error rate is important, the methods of Voss (1988, 1999) or Wu and Wang's (2006) may be used, since these methods control EER in the strong sense; see also Voss and Wang (2006).

In the end, we emphasize that our theoretical result (no test that performs well for all configurations of active and inactive effects) is valid in the absence of any prior information (including sparsity) concerning active and inactive effects. As noted at the first of this section, some methods can be recommended if one has a good idea about how many active effects there are.

Acknowledgements

The authors wish to thank the Research Council of Shiraz University. They also would like to thank the referees for their valuable comments.

References

Aboukalam, M. A. F., and Al-Shiha, A. A. (2001), A robust analysis for unreplicated factorial experiments. *Computational Statistics & Data Analysis*, **36**, 31-46.

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

Al-Shiha, A. A., and Yang, Shie-Shien (2000), Critical values and some properties of a new test statistic for analyzing unreplicated factorial experiments. *Biometrical Journal*, **42**, 605-616.

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

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

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.

Birnbaum, A. (1961), A multi-decision procedure related to the analysis of single degrees of freedom. *Annals of the Institute of Statistical Mathematics*, **12**, 227-236.

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

Chen, Y. (2003), *On the analysis of unreplicated factorial designs.* (PH.D. thesis submitted to the Department of Statistics of the University of Dortmund, Dortmund, March 2003).

Chen, Y., and Kunert, J. (2004), A new quantitative method for analysing unreplicated factorial designs. *Biometical Journal*, **46**(1), 125-140.

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

Dong, F. (1993), On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, **3**(1), 209-217.

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

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

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 and Analysis (Edited by C. L. Mallows)*. John Wiley, New York.

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

Kharrati-Koapei, M., and Sadooghi-Alvandi, S. M. (2007), A new method for testing interaction in unreplicated two-way analysis of variance. *Communications in Statistics-Theory and Methods*, **36**, 2787-2804.

Kharrati-Kopaei, M., and Shenavari, Z. (2023), Analyzing unreplicated two-level factorial designs by combining multiple tests. *Communications in Statistics-Theory and Methods*, 1-16.

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 and Data Analysis*, **14**, 135-148.

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

Miller, A. (2005), The analysis of unreplicated factorial experiments using all possible comparisons. *Technometrics*, **47**(1), 51-63.

Montgomery, D. C. (2001), *Design and analysis of experiments. Fifth edition*. John Wiley & Sons, Inc.

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 Math*, **21B**, 57-98.

Venter, J. H., and Steel, S. J. (1996), A hypothesis-testing approach toward identifying active contrasts. *Technometrics*, **38**(2), 161-169.

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

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

Voss, D. T., and Wang, W. (2006), On adaptive testing in orthogonal saturated designs. *Statistica Sinica*, **16**, 227-234.

Westfall, P. H., and Young, S. S. (1993), *Resampling-based multiple testing, examples and methods for p-value adjustment.* John Wiley & Sons, INC.

Wu, S. S., and Wang, W. (2007), Step-up simultaneous tests for identifying active effects in orthogonal saturated designs. *Annals of Statistics*, **35**, 449-463.

Ye, K. Q., Hamada, M., and Wu, C. F. J. (2001), A step-down Lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology*, **33**(2), 140-152.

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.