Considering the risk of ignoring active factors in industrial experiments

Pere Grima, Lourdes Rodero and Xavier Tort-Martorell

Department of Statistics and Operational Research
Universitat Politècnica de Catalunya - BarcelonaTech, Spain

Introduction

When interpreting the results of a statistical analysis like a comparison of treatments or factorial design, emphasis is generally placed on not making a mistake when declaring certainty or near (95% certainty) that a particular factor is active. In other words, the interpretation and analysis itself focus on minimizing the risk of type I error. This is the standard and appropriate approach to many situations. However, industry technicians often assert that they would rather commit the error of ignoring a factor that in reality affects a response rather than ignoring a factor that is in reality affecting the response. In statistical terms, they prefer to minimize type II errors.

This article draws attention to the importance of considering type II error in Design of Experiments contexts and shows, through two examples, an easy way to do that.

Why type II error is highly relevant to business and industry

Unsurprisingly, type I error causes more concern than type II does in scientific environments, while often the opposite occurs in the industrial and business worlds. Nor is it astonishing that such concern for type I error is the standard, since it was only until recently that experimental design remained concentrated mostly in scientific fields.

In the scientific world, when it is stated that factor (*A)* affects the response (*Y*) in a particular way, we want to be as certain as possible that this is true. The fact that another factor (*B*) also affects it and has gone unnoticed is relevant; but it is not an error that nullifies either the veracity or the utility of the investigation. On the contrary, affirming that *A* affects *Y* when this is not true is rightly considered an unacceptable error.

This is different in the industrial and business worlds, where erroneously concluding that *A* affects *Y* can lead to problems like unnecessarily controlling for *A*, although practice and experience will likely reveal the error sooner or later. A much bigger problem is not detecting that *B* affects *Y*, as it can cause multiple difficulties such as the variation of *Y* increasing as a result of failing to control *B* or cost overruns from trying to optimize *Y* without accounting for *B*'s effect. What makes this error even worse is that it will be very difficult to detect such an effect without conducting a new study on whether or not *B* has a relevant effect on *Y* – something that will rarely happen immediately after the first study.

However, when analyzing which effects should be considered active, attention is usually paid only to the probability of committing the first type of error, which is generally referred to as type I error and is usually set at 5%. Meanwhile, the probability of the second – the so-called type II error – is ignored and often turns out to be much greater than what one could consider reasonable. De León *et al*. (2006) showed that, in DEO contexts, this error probability is usually greater than the experimenter suspects.

Proposed methodology

When we analyze the significance of the estimated effects after conducting a factorial experiment, what we are trying to do is accurately divide them into two groups: inert and active. However, in practice, this is impossible for two reasons. First, just by chance, we can obtain effect estimates that are greater than, say, two times their standard errors when the true effect is actually zero. Second, we can quite frequently obtain effect estimates that are smaller than two times their standard error when their real effect is slightly greater.

Bearing this in mind, particularly in the abovementioned business and industrial context, it makes sense to define the minimum real value of the effects that we want to detect, that is, those we want to be considered significant. In other words, the idea is to set a value such that any effect that exceeds it has a low probability of being ignored. Following De León *et al*. (2006), we shall refer to this as the minimum effect size of interest ($MESI$). In industrial contexts, setting this value is not only relatively simple, but we believe that it is always advisable, as it forces the experimenter to think about the variability (standard error) of the effects and relate it to the variation of the system, the choice of factor levels, and the number of runs.

The following proposal is based on a more formal one presented by Grima *et al*. (2020), briefly commented on in the section titled "A more formal alternative to step 3", and in a suggestion from Professor Alberto Luceño (Universidad de Cantabria, Spain) in a personal communication to one of the authors. It consists of three simple steps:

1. Set the $MESI$ value.

   As stated, MESI should be set at a value that, if exceeded, is considered by the experimenter to indicate relevant effects, meaning they are affecting the response in an important way. Therefore, we want a low probability of considering the estimated effect of such variables to be non-significant.

2. Estimate the effect's standard error.

   A common way to estimate this standard error is through Lenth's pseudo standard error (PSE; Lenth, 1989). The PSE is based on the fact that if $X \sim \mathrm{N}(0, \sigma^2)$, the median of $|X|$ equals 0.6745σ and therefore $1.5 \cdot \mathrm{median}|X| = 1.01\sigma \cong \sigma$. Assuming that e_i $(i = 1, \ldots, n)$ are the values of the effects of interest and that their estimators $\hat{e}_i$ are distributed according to $\mathrm{N}(e_i, \sigma_e^2)$, Lenth defines $s_0 = 1.5 \cdot \mathrm{median}|\hat{e}_\mathrm{i}|$ and calculates a new median by excluding the estimated effects with $|\hat{e}_i| > 2.5 s_0$. In doing so, he expects to exclude the effects with $e > 0$ and use the others to calculate:

$$PSE = 1.5 \cdot \underset{|\hat{e}_i| < 2.5 s_0}{\mathrm{median}} |\hat{e}_i|$$

3. Calculate 95% confidence intervals for the effects and assess the differences between interval limits and the $MESI$, and between the interval limits and 0.

It is well known (Loughin, 1998; Ye and Hamada, 2000; Fontdecaba et al., 2015) that Lenth's recommended degrees of freedom for estimating the PSE (still used by very popular statistical software packages) lead to lower type I error probabilities than the intended value ($\alpha = 0.05$), with the undesirable consequence of a higher type II error probability. Therefore, to calculate 95% confidence intervals for the estimated effects, we propose using the proposal of Fontecaba *et al*. (2015): using a t value of 2 to calculate 95% confidence intervals for both 8- and 16-run designs. Not only does this provide better results than those in Ye and Hamada's approach, but it is simpler.

So, we calculate the intervals as effect $\pm\ 2{\cdot}PSE$ and we plot these interval for all the effects together with 0 and the MESI. It is then easy to classify the effects into three groups:

- Intervals that do not contain 0. Clearly significant effects.
- Intervals that contain 0 but not the MESI. Clearly non-significant effects.
- Intervals that contain 0 and the MESI. Then, the effect should be considered significant if 0 is closer than the MESI to the center of the interval and non-significant otherwise. Alternatively, the effect should be studied in more detail

Let us turn now to some simple examples of applying this methodology, taking the data from DOE examples of very well-known texts.

Example 1: A 2^3 design

Based on data published by Prat and Tort-Martorell (1990), Box *et al*. (2005, p. 194) present a 2^3 design carried out in a pet food factory. Several responses are analyzed, but we consider only the amount of product obtained (yield). The factors considered are:

Factors	Levels $-$	Levels $+$
A: Conditioning temperature	80% of max.	Max
B: Flow	80% of max.	Max
C: Compression zone	2	2.5

Table 1 shows the design matrix together with the results obtained (left) and the values of the effects (right).

Table 1: Design matrix and effects obtained in the example by Box et al. (2005).

Factors			y
A	B	C	
-1	-1	-1	83
1	-1	-1	85
-1	1	-1	99
1	1	-1	102
-1	-1	1	59
1	-1	1	75
-1	1	1	80
1	1	1	73

Effects	
A:	3.5
B:	13.0
C:	-20.5
AB:	-5.5
AC:	1.0
BC:	-3.5
ABC:	-6.0

Looking at Figure 1, it is clear that factor C should be considered significant, although doubts may arise regarding the possible influence of factor B. The line on the Pareto chart (calculated using t = 2) indicates the boundary between the effects that should be considered significant and those that should not.

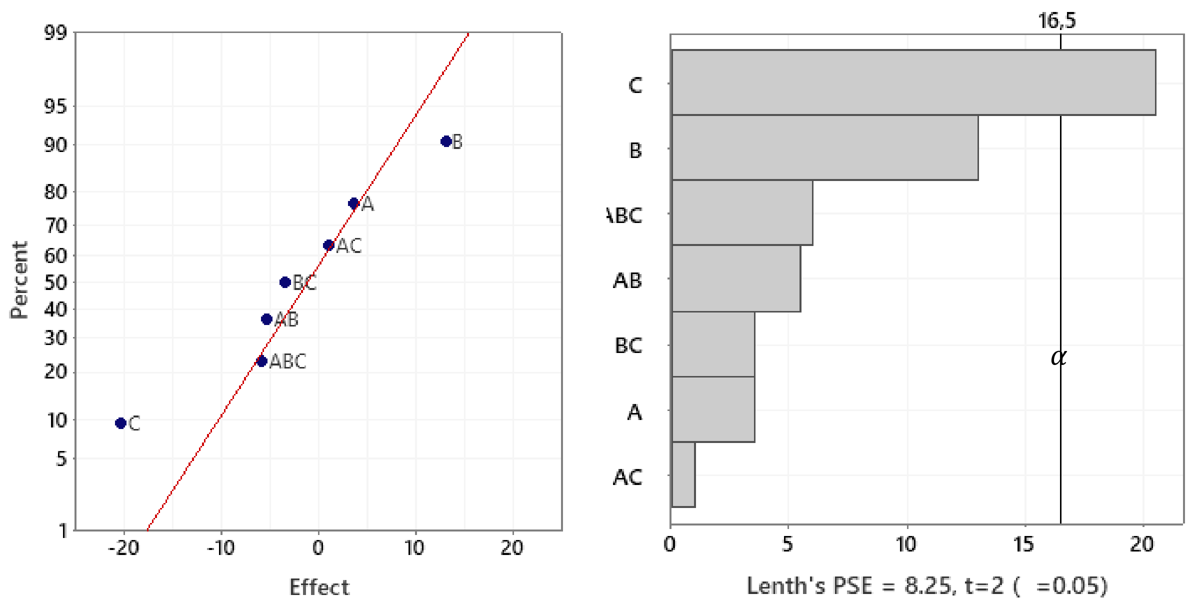

Figure 1: NPP (left) and Pareto chart (right) representations of the effects in the example by Box et al. (2005)

Following our proposal, we have:

1. Set the MESI value
   In this case, let us suppose that a 20-unit change in yield is a relevant change ($MESI = 20$)

2. Estimate the effect's standard error.
   In this case, $median|e_i| = 5.5$. Thus, $s_0 = 1.5 \cdot 5.5 = 8.25$. Now, we should exclude the effects with $|\hat{e}_i| > 2.5s_0 = 20.625$; but, since there are none, we have that PSE = 8.25.

3. Calculate 95% confidence intervals for the effects and assess the differences between interval limits and the $MESI$, and between the interval limits and 0.
   As stated, once the PSE has been obtained, these intervals are very easy to compute by simply calculating: effect ± 2·PSE.

After plotting the MESI and the zero references, and the obtained intervals, we obtain the result in Figure 2, which clearly shows that C is a significant effect. It also shows, that for B, the MESI value (20) is far more probable than 0. Therefore, and this is the point, that it should be considered as significant.

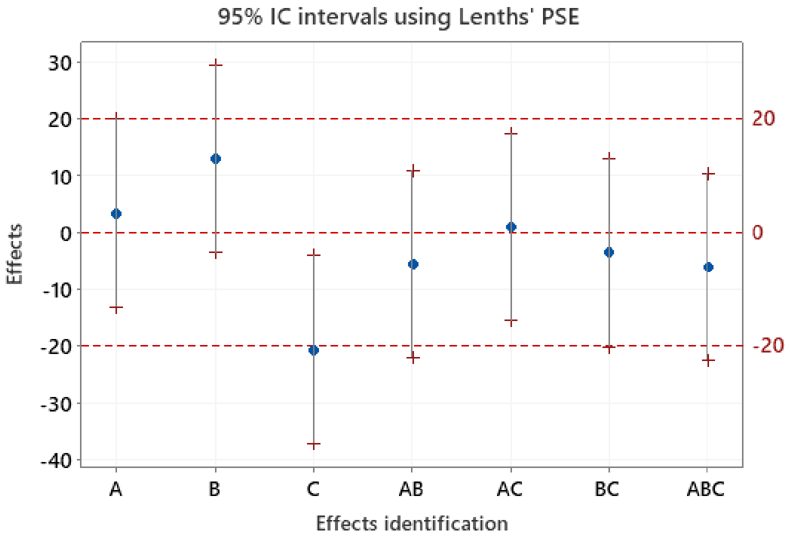

Figure 2: 95% CI for the effects, including 0 and MESI references for the 2^3 example.

Example 2: A 2^4 design

Based on an article by G.H. Bell *et al*. (2006), Montgomery (2013, p. 279) presents a 2^4 design to analyze whether certain changes in the conditions of a credit card increase the number of affirmative responses. The factors considered are:

Factor	Level	
	–	+
A: Annual fee	Current	Lower
B: Account-opening fee	No	Yes
C: Initial interest rate	Current	Lower
D: Long-term interest rate	Low	High

For each combination of factor values, 7500 letters were sent with the corresponding offer. On the left side of Table 2 is the design matrix along with the responses obtained. On the right, we have the estimated effects.

Table 2: Design matrix and effects obtained in the Montgomery example (2013)

Factors				
A	B	C	D	y
-1	-1	-1	-1	184
1	-1	-1	-1	252
-1	1	-1	-1	162
1	1	-1	-1	172
-1	-1	1	-1	187
1	-1	1	-1	254
-1	1	1	-1	174
1	1	1	-1	183
-1	-1	-1	1	138
1	-1	-1	1	168
-1	1	-1	1	127
1	1	-1	1	140
-1	-1	1	1	172
1	-1	1	1	219
-1	1	1	1	153
1	1	1	1	152

Effects	
A:	30.37
B:	-38.87
C:	18.875
D:	-37.37
AB:	-22.63
AC:	0.125
AD:	-8.125
BC:	-3.625
BD:	7.625
CD:	11.875
ABC:	-3.875
ABD:	6.375
ACD:	0.625
BCD:	-8.125
$ABCD$:	-3.875

Figure 3 presents the NPP and the Pareto diagram of the effects. The main effects A, B, and D are clearly significant; AB generates doubts; and so does C, to a lesser extent.

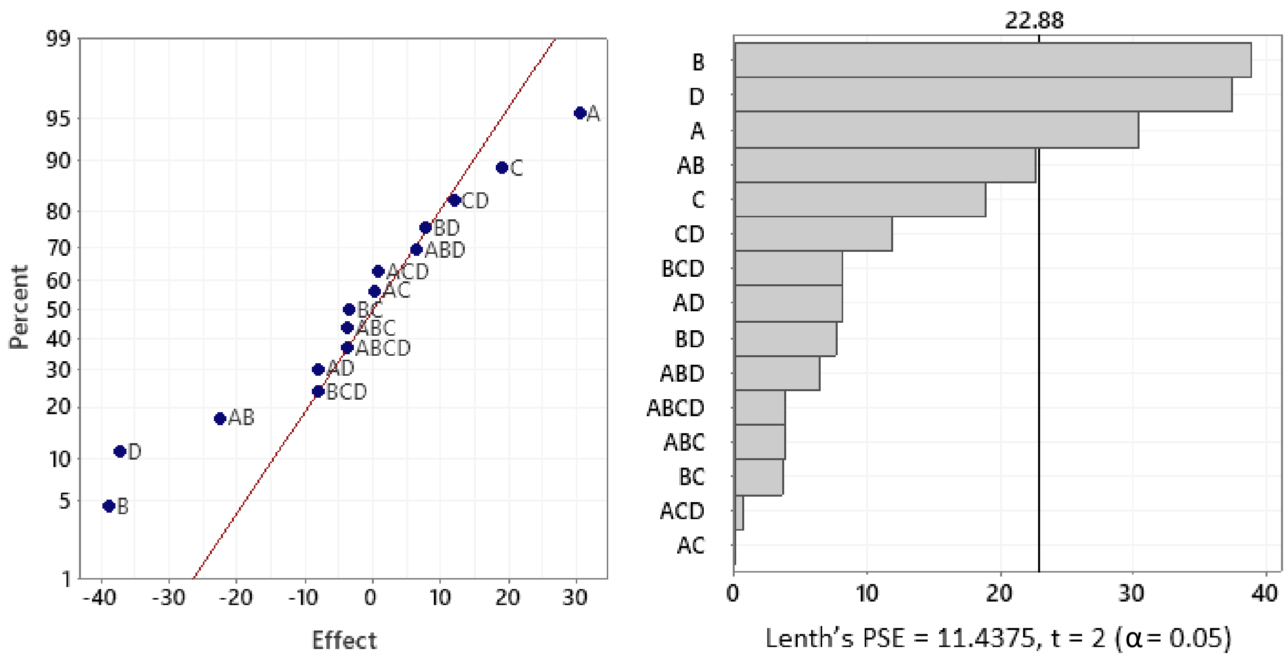

Figure 3: NPP (left) and Pareto chart (right) representations of the effects in the example by Montgomery (2013)

Now, let us clarify the situation using the three-step procedure:

1. Set the MESI value
   Let us suppose we do not want an effect to go unnoticed when its real value is greater than 30, a rather large value in light of the results obtained.

2. Estimate the effect's standard error
   As in the previous example, we use the PSE, which here is PSE = 11.4375. In this case, we could alternately use the estimation of the five higher order interactions.

3. Calculate 95% confidence intervals for the effects and assess the differences between interval limits and the $MESI$, and between the interval limits and 0.

Again, we are using Fontdecaba's approximation to obtain that the 95% CI is: effect ± 2(11.4375)

Thus, we obtain Figure 4, which indicates that both the AB interaction and factor C also have a high probability of relevantly affecting the response.

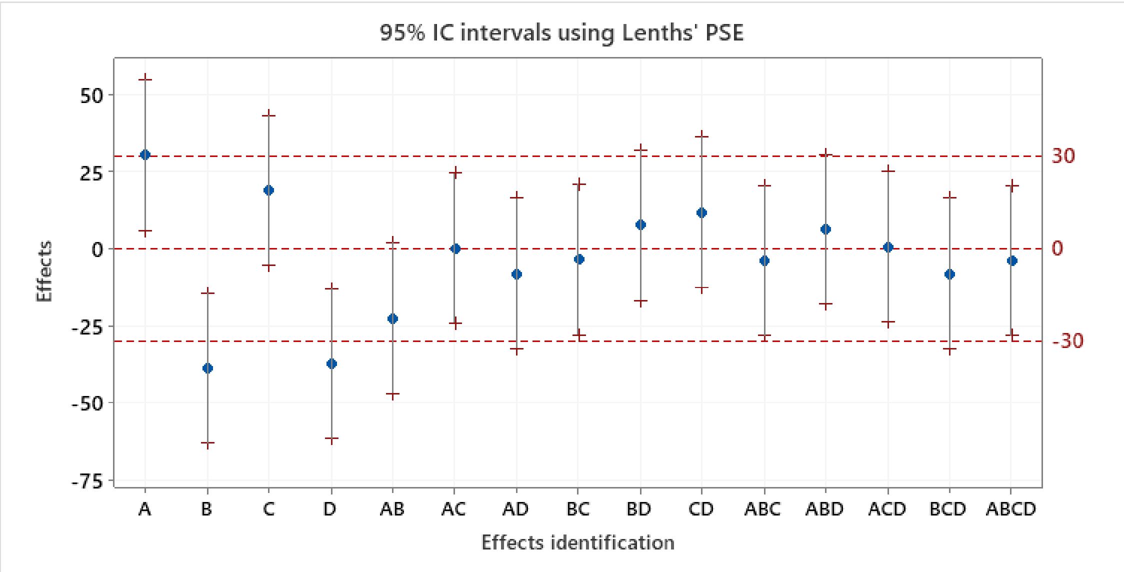

Figure 4: 95% CI for the effects, including 0 and the MESI references for the 2^4 example.

A more formal alternative to step 3

A more formal way to solve the problem is to set the probability that an effect is considered non-significant when its real value is equal to the MESI. This is what we call the type II error probability, or β risk. By fixing it at a desired level, we can establish a critical value that complements the usual criterion of α = 0.05, thus allowing us to represent two boundary lines in the Pareto diagram of the effects: the usual one, based on α = 0.05; and a new one based on the β probality. Then, the effects can be classified into three groups:

- The effects that exceed the α line, which are almost surely active.

- The effects below the β line, which should be considered non-active.
- The effects between the two lines should be considered as active or in any case as worth further study.

Figure 5 shows the Pareto diagrams of the two analyzed examples with the two boundary lines. In both cases, σ_e was estimated using the standard deviation of the effects that are considered not significant, for which we applied the Lenth method. This gives us $s_e = 6.59$ with 6 degrees of freedom for the 2^3 design and $s_e = 10.34$ with 12 degrees of freedom for the 2^4. The new critical values are those shown in the diagram. Details on their calculation can be found in Grima et al. (2020).

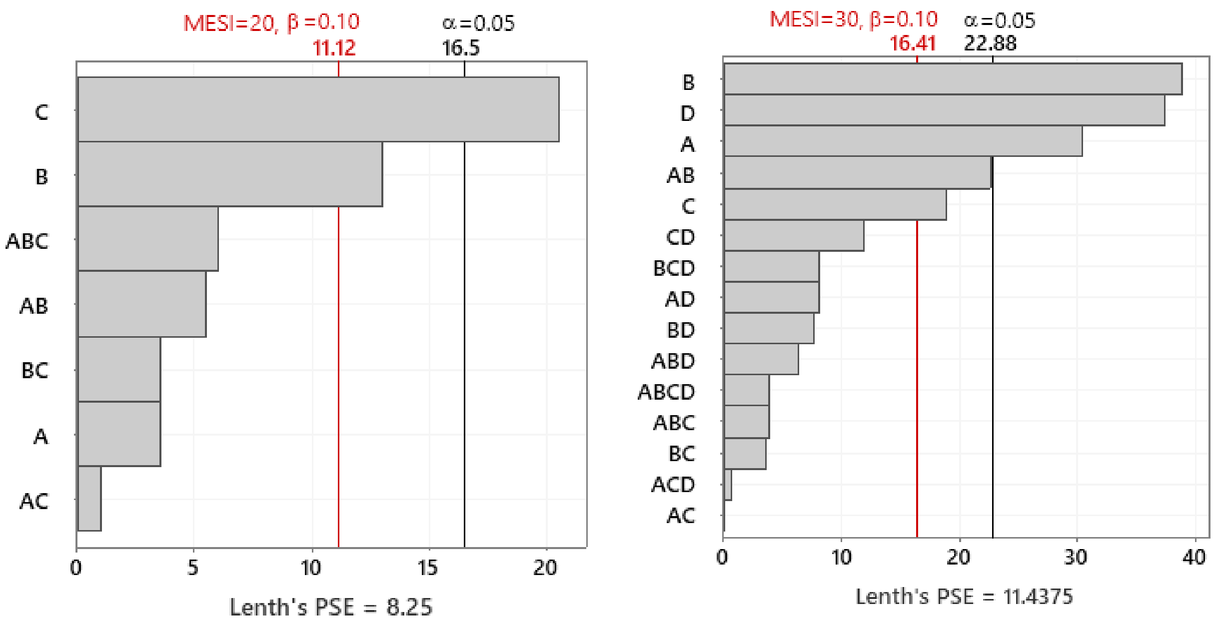

Figure 5: Pareto diagrams of the effects for the two examples with two critical lines: the usual one (based on a type I error probability of α = 0.05); and another based on MESI and a type II error probability of (β = 0.10)

Conclusion

In industrial experiments, it is usually interesting to minimize the risk of overlooking active effects. One way to accomplish this is to establish a MESI, a minimal effect size of interest, and based on it use either 95% confidence intervals for the effects or the critical value for a certain type II error probability to classify them into active, inert or worth to be further studied. Apart from the utility of identifying active factors that would have otherwise been overlooked, focusing on the $MESI$ indeed provokes a worthwhile and interesting reflection on the part of the experimenter.

References

Bell, G.H., J. Ledolter, and A.J. Swersey. 2006. *Experimental design on the front lines of marketing: Testing new ideas to increase direct mail sales*, International Journal of Research in Marketing, 23(3): 309–319. https://doi.org/10.1016/j.ijresmar.2006.05.002

Box, G. E. P., Hunter, J. S., Hunter, W. G. (2005). Statistics for experimenters: Design, innovation, and discovery, 2nd edition. NJ: Wiley-Interscience

De León G, Grima P, Tort-Martorell X. Selecting Significant Effects in Factorial Designs Taking Type II Errors into Account. *Quality and Reliability Engineering International* 2006; 22(7):803-810. https://doi.org/10.1002/qre.729

Fontdecaba, S., P. Grima and X. Tort-Martorell. 2015. Proposal of a single critical value for the Lenth method. *Quality Technology and Quantitative Management,* 12(1):41-51. https://doi.org/10.1080/16843703.2015.11673365

Grima P, Rodero L, Tort-Martorell X. Selecting relevant effects in factorial designs. *Quality and Reliability Engineering International* 2020; 36(7):2370-2378. https://doi.org/10.1002/qre.2702

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

Loughin, T. M. 1998. Calibration of the Length test for unreplicated factorial designs. *Journal of Quality Technology*, 30(2): 171-175. https://doi.org/10.1080/00224065.1998.11979836

Montgomery, D.C. (2013) *Design and analysis of experiments*, Willey, Singapore

Ye, K.Q. and M. Hamada. 2000. Critical values of the Lenth method for unreplicated factorial designs. *Journal of Quality Technology,* 32(1):57-66. https://doi.org/10.1080/00224065.2000.11979971