# Avoiding Mean Square Error Bias in Designed Experiments

James A. Colton

**To avoid a biased MSE term, statistically valid model-reduction techniques should be used and the experiment should be properly randomized.**

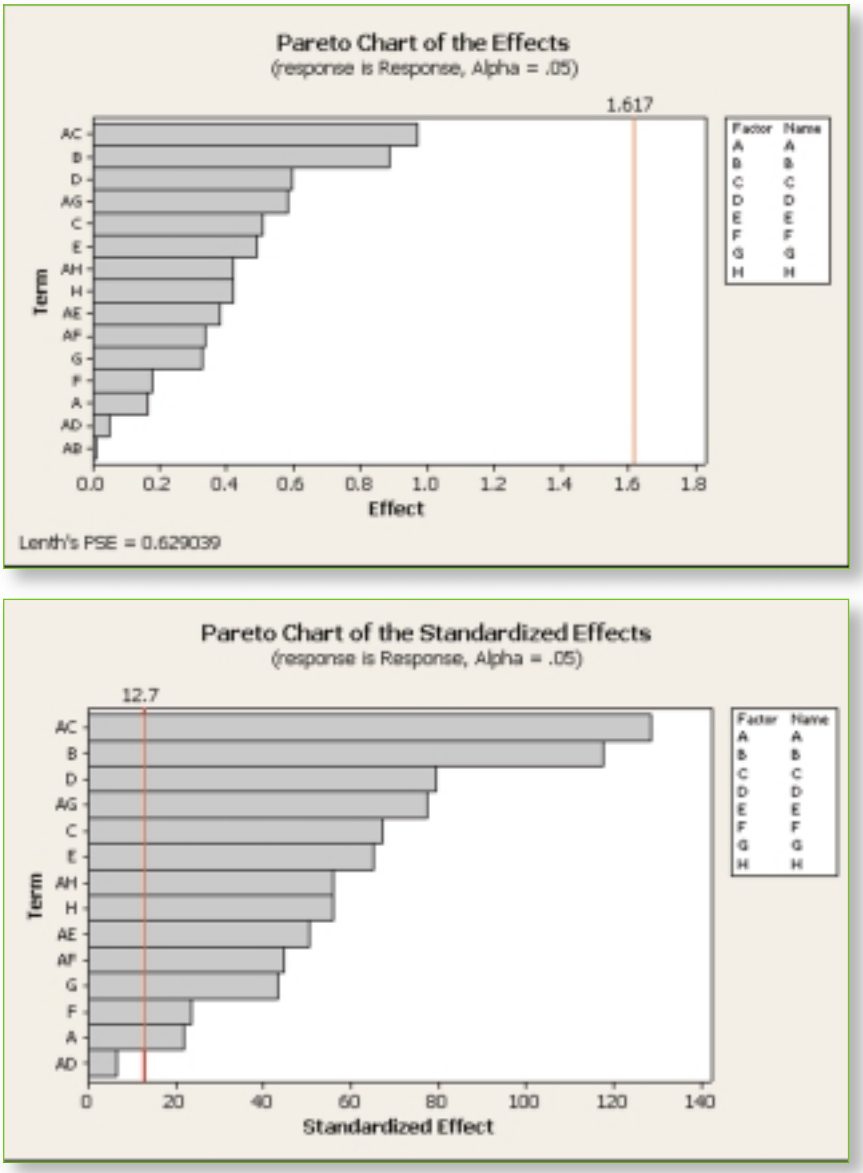


**Chart 1:** *Pareto chart of effects for an 8-factor,16-run experiment before (top) and after (bottom) eliminating the smallest effect. Any effect extending beyond the vertical red line is significant at the 0.05 α-level.*

The error in a designed experiment, $\sigma^2$, is the natural variation in the response when one of the experimental combinations is replicated. One important challenge in a designed experiment is obtaining an unbiased estimate of $\sigma^2$. Too often, experimenters do not realize the impact that data collection and analysis assumptions have on the estimate of $\sigma^2$. If the estimate is biased, tests of the effects in the analysis will be adversely affected. This article discusses the causes and consequences of bias in the mean square error (MSE) term and provides suggestions for detecting and correcting MSE bias.

## Introduction

A replicated experiment runs each factor combination more than one time and calculates an error estimate ($\sigma^2$) at each experimental combination. These estimates are pooled to obtain an overall estimate of $\sigma^2$. An unreplicated experiment obtains an estimate by assuming that higher-order interactions are "noise" or by using a modern technique such as Lenth's pseudo standard error [1].

The ANOVA table refers to the estimate of $\sigma^2$ as the mean square error (MSE), which is based on the following:

1. replication (pure) error
2. terms removed from the model (lack of fit error)
3. a combination of (1) and (2).

## Causes of MSE Bias

In this article, the term bias refers to the difference between the estimated value of MSE and the expected or true value. Positive bias occurs when the estimated value is higher than the true value and negative bias occurs when the estimated value is lower than the true value.

Bias can occur in the pure-error

component or the lack-of-fit component of MSE. Typically, pure-error bias is the result of unsound data-collection procedures while lack-of-fit (LOF) error bias is the result of an inappropriate model-reduction approach.

***Pure-error bias:*** Pure-error bias arises when the variation between replicates does not represent the natural variation in the process. If replicates are collected at the same time and/or under the same conditions, they will likely be exposed to the same or similar environmental conditions, raw materials, operators, and machinery. As a result, you can expect a negative bias in the pure-error component when non-randomized replicates (sometimes referred to as repeats) are incorrectly treated as replicates. You also can expect a negative bias in the pure-error component when multiple measurements from the same part are incorrectly treated as replicates.

A positive bias in the pure-error component can occur when one replicate of an experiment is collected under different conditions than a second replicate, and a blocking variable is not used to account for the differences between replicates. In this case, an additional source of variation is included in the pure-error component, resulting in a pure-error component that has a positive bias.

***Lack-of-fit error bias:*** LOF-error bias arises when terms are removed from the model incorrectly. A negative bias in LOF error can occur when insignificant terms in an orthogonal experimental design (a design in which the effects of any factor are balanced across the effects of the other factors) are removed from the model one at a time, starting with the term with the largest p-value, or equivalently, the smallest absolute effect. With this approach, the analyst is hand-picking the smallest possible measure of $\sigma^2$, placing it in the error, and using it as an estimate of $\sigma^2$ in the next stage of the analysis. In this situation, you can expect a negative bias in the LOF component, which can be especially dangerous if the LOF error is the only component of variation in the MSE (such as in unreplicated designs). This danger can be demonstrated via simulation.

Consider an 8-factor, 16-run resolution IV designed experiment with a random response obtained from a data simulator. Chart 1 shows the Pareto-effects chart before (top) and after (bottom) removing the term with the smallest absolute effect. With all terms in the model, the chart correctly indicates that none of the effects are significant. After the term with the smallest absolute effect is

**Table 1:** *The standard error (SE) for each coefficient in the model is based on the MSE. The magnitude of each coefficient is compared to the SE to generate a t statistic. The t statistic is converted into a p-value that provides a measure of statistical significance.*

| MSE | SE | T-Statistics | P-Values | Reject $H_0$ | Type I Error |
|---|---|---|---|---|---|
| Negative Bias → | Negative Bias → | Farther from 0 → | Closer to 0 → | More Often → | Increased |

| MSE | SE | T-Statistics | P-Values | Reject $H_0$ | Type II Error |
|---|---|---|---|---|---|
| Positive Bias → | Positive Bias → | Closer to 0 → | Closer to 1 → | Less Often → | Increased |

**Table 2:** *Indicators of MSE bias*

| Indicator | Outcome |
|---|---|
| All P-values for terms in the full model are close to 1 | MSE may have a positive bias |
| All P-values for interaction terms in the full model are close to 0 | MSE may have a negative bias |
| P-value for the lack-of-fit F-test (see Appendix) is close to 0 | Pure error may have a negative bias and/or lack-of-fit error has a positive bias |
| P-value for the lack-of-fit F-test (see Appendix) is close to 1 | Pure error may have a positive bias and/or lack-of-fit error has a negative bias |

**Table 3:** *Remedies for MSE bias*

| Form of Bias | Possible Cause | Remedy |
|---|---|---|
| LOF error has positive bias. | Significant effects were incorrectly removed from the model and are in the error term. | Refit the model, being careful not to eliminate significant effects. If the design is non-orthogonal, eliminate the terms one at a time. |
| LOF error has negative bias. | Orthogonal terms were eliminated from an unreplicated design one at a time, starting with the term that has the largest p-value. | Refit the model using Lenth's approach (default in MINITAB) to identify significant effects. |
| Pure error has positive bias. | The first replicate of the experiment was collected under different environmental conditions than the second replicate. | Block on replicates. The blocking variable removes the effect of the blocking condition from the pure error. |
| Pure error has negative bias. | Replicates have not been randomized or replicates are actually repeated measurements of the same part. | Average the replicates at each factor combination. The replicates are no longer used to estimate MSE and there is no longer a pure-error term. |

**Table 4:** *Two replicates of a four-factor, full-factorial design. Three- and four-factor interactions are assumed to be noise.*

Fractional Factorial Fit: Response versus A, B, C, D
Estimated Effects and Coefficients for Response (coded units)

| Term | Effect | Coef | SE Coef | T | P |
|---|---|---|---|---|---|
| Constant | | 0.0875 | 0.1478 | 0.59 | 0.560 |
| A | 0.1318 | 0.0659 | 0.1478 | 0.45 | 0.660 |
| B | -0.1745 | -0.0872 | 0.1478 | -0.59 | 0.561 |
| C | -0.8701 | -0.4350 | 0.1478 | -2.94 | 0.008 |
| D | -0.2787 | -0.1394 | 0.1478 | -0.94 | 0.356 |
| A*B | -0.5209 | -0.2605 | 0.1478 | -1.76 | 0.093 |
| A*C | -0.0783 | -0.0392 | 0.1478 | -0.26 | 0.794 |
| A*D | 0.1932 | 0.0966 | 0.1478 | 0.65 | 0.520 |
| B*C | -0.5128 | -0.2564 | 0.1478 | -1.73 | 0.097 |
| B*D | 0.0300 | 0.0150 | 0.1478 | 0.10 | 0.920 |
| C*D | 0.1111 | 0.0555 | 0.1478 | 0.38 | 0.711 |

Analysis of Variance for Response (coded units)

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Main Effects | 4 | 7.060 | 7.060 | 1.7651 | 2.52 | 0.071 |
| 2-Way Interactions | 6 | 4.728 | 4.728 | 0.7880 | 1.13 | 0.381 |
| Residual Error | 21 | 14.683 | 14.683 | 0.6992 | | |
| Lack of Fit | 5 | 7.400 | 7.400 | 1.4800 | 3.25 | 0.032 |
| Pure Error | 16 | 7.283 | 7.283 | 0.4552 | | |
| Total | 31 | 26.471 | | | | |

removed, the remaining 14 effects are compared to it and 13 are found to be significant, which results in 13 Type I errors (a Type I error occurs when an effect is incorrectly deemed significant).

Another inappropriate model-reduction technique can result in a positive bias in the LOF-error component. This occurs when two or more non-orthogonal terms are eliminated from the model together in the same step. One of these terms actually may be significant if left in the model without the other term. The MSE is then higher than expected because the error for this term is included in the MSE when it should not be. This commonly happens in response surface designs when removing quadratic terms that are non-orthogonal.

## Consequences of bias

If the MSE is biased, then all the tests for significant effects either have an inflated Type I error or an inflated Type II error (a Type II error occurs when an effect is incorrectly deemed not significant). An MSE with a negative bias results in an inflated Type I error rate in the tests for significant effects (Table 1). As a result, process settings may be unnecessarily adjusted and future experiments may contain unimportant factors, which may increase experiment costs. An MSE with a positive bias results in an inflated Type II error rate in the test for significant effects. As a result, the analysis may exclude factors or interactions that influence the response.

**If the mean square error is biased, then all the tests for significant effects either have an inflated Type I error or an inflated Type II error.**

## Signs and remedies

Listed in Table 2 are four indicators that should raise suspicions that the MSE is biased. Specific cutoffs for p-values are not given in Table 2 because formal tests for most forms of MSE bias have not been developed and, in some cases, depend on how many terms are in the model. Future research into formal p-value cut-offs would be beneficial. If one of the indicators of MSE bias from Table 2 is present, Table 3 can be used to identify the possible cause and to recommend a remedy.

## Summary

It is important to carefully inspect the data collection and model reduction process for signs of bias in MSE before the results are used to reach any conclusions. To avoid a biased MSE term, statistically valid model-reduction techniques should be used and the experiment should be properly randomized.

## Appendix

The lack-of-fit test requires both a pure-error component and a LOF-error component. The null hypothesis states that the terms removed from the model to form the LOF error are null effects (effects that do not influence the response and can be used to estimate noise). The alternative states that at least one term removed from the model is an active effect (an effect that influences the response and should not be used to estimate noise), which implies that the MSE has a positive bias. In MINITAB, the lack-of-fit test is printed by default when both a pure-error and LOF-error component exist. In the example in Table 4, the F-test for the lack-of-fit error component has a p-value of 0.032, which might indicate a bias in the MSE.

## References

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

*James A. Colton is a Technical Training Specialist at Minitab Inc. He may be contacted at sceditor@scimag.com.*