Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: www.tandfonline.com/journals/lqen20

On scoping a test that addresses the wrong objective

Thomas H. Johnson, Rebecca M. Medlin, Laura J. Freeman & James R. Simpson

To cite this article: Thomas H. Johnson, Rebecca M. Medlin, Laura J. Freeman & James R. Simpson (2019) On scoping a test that addresses the wrong objective, Quality Engineering, 31:2, 230-239, DOI: 10.1080/08982112.2018.1479035

To link to this article: https://doi.org/10.1080/08982112.2018.1479035

Published online: 13 Nov 2018.

QUALITY ENGINEERING
2019, VOL. 31, NO. 2, 230–239
https://doi.org/10.1080/08982112.2018.1479035

Taylor & Francis
Taylor & Francis Group

On scoping a test that addresses the wrong objective

Thomas H. Johnson[a], Rebecca M. Medlin[a], Laura J. Freeman[a], and James R. Simpson[b]

[a]Operational Evaluation Division, Institute for Defense Analyses, Alexandria, Virginia; [b]JK Analytics, LLC, Niceville, Florida

ABSTRACT

The Department of Defense test and evaluation community uses power as a key metric for sizing test designs. Power depends on many elements of the design, including the selection of response variables, factors and levels, model formulation, and sample size. The experimental objectives are expressed as hypothesis tests, and power reflects the risk associated with correctly assessing those objectives. Statistical literature refers to a different, yet equally important, type of error that is committed by giving the right answer to the wrong question. If a test design is adequately scoped to address an irrelevant objective, one could say that a Type III error occurs. In this paper, we focus on a specific Type III error that on some occasions test planners commit to reduce test size and resources. We provide a case study example that shows how reparameterizing a factor space from fewer factors with more levels per factor to a space that has more factors with fewer levels per factor fundamentally changes the hypothesis tests, and hence may no longer be aligned with the original objectives of the experiment. Despite the perceived increase in power and decrease in test resources that comes from this reparameterization, we conclude, it is not a prudent way to gain test efficiency. Through the case study example, we highlight the information that is lost in this decision and its implications on test objectives.

KEYWORDS

design of experiments; sample size determination; hypothesis testing; statistical power; Type III error; unified effect size

Introduction

Design of experiments is used across a variety of fields to aid in the planning, execution, and analysis of an experiment. In the planning phase, critical questions about the system under test are identified and the experimental objectives are set. These questions and objectives guide the development of the response variables, factors, and levels (Freeman et al., 2013). Recent Department of Defense policy has emphasized the importance of using design of experiments in the operational testing of all military systems (Johnson et al., 2012).

Equally important in the planning phase is the evaluation of the experimental design. An assortment of measures are available to assess the adequacy of a design, prior to data collection. Hahn, Meeker, and Feder (1976) call these "measures of precision," which include the standard error of predicted mean responses, standard error of coefficients, correlations metrics, and optimality criteria values. Measures of precision are affected by many aspects of the plan, including the choice of factors and levels, the assumed model form, the combination of factor settings from run to run, and the total number of runs in the experiment.

An additional and widely used measure of precision, especially in the Department of Defense test and evaluation community, is power. When the objective of the experiment, such as determining whether a new weapon system is better than an old system, is expressed as a hypothesis test, power informs the risk associated with correctly assessing that objective. Because power increases with sample size, it is a useful metric for determining test length and test resourcing.

An adequate experiment requires sufficient power, but more importantly it requires the hypothesis tests to accurately reflect the test objectives. If an experiment provides adequate power, but addresses the wrong objective, we might say an error is committed. Kimball (1957) refers to this error as an error of the third kind, stating, "[A Type III error] is the error committed by giving the right answer to the wrong question."

Problem statement

In this paper, we focus on a Type III error that test planners might commit in an attempt to reduce test size and test resources. We provide an example that

CONTACT Rebecca M. Medlin rmedlin@ida.org Institute for Defense Analyses, Alexandria, 22311-1882, VA.

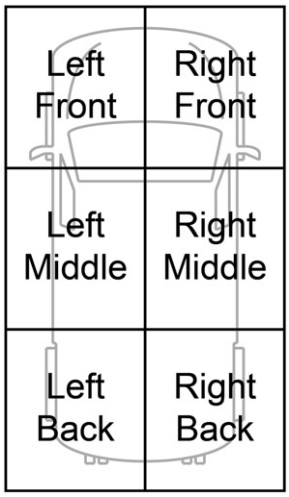

Figure 1. Vehicle under-body detonation locations.

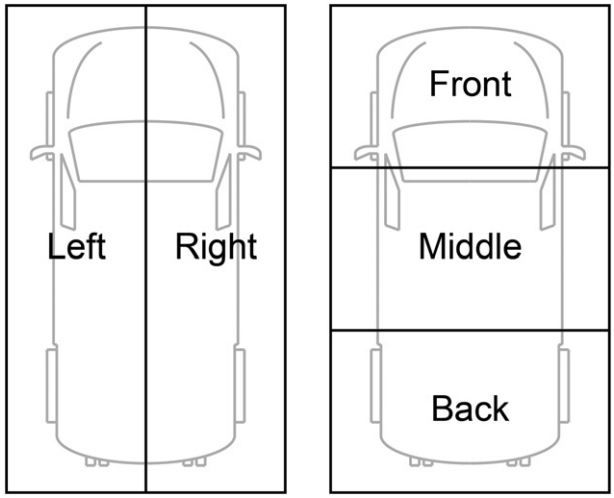

Figure 2. Reparameterization of the six-level Location factor into a two-level Side factor and a three-level Position factor.

shows how reparameterizing a factor space from fewer factors with more levels per factor to a space that has more factors with fewer levels per factor fundamentally changes the hypothesis tests, and hence may no longer be aligned with the original objectives of the experiment.

Consider, as a case study, an experiment that plans to characterize a vehicle's vulnerability against a particular type of mine. The test program has a limited number of vehicles and mines at their disposal to run a series of destructive tests to characterize the vehicle's vulnerability.

The response variable is measured as the static deformation of the vehicle's under-body armor plate after interaction with the blast wave and ejecta from the buried charge. In other words, deformation is a direct measurement of the vehicle's armor shape change with respect to a reference point.

The engineering team surmises that the non-uniform placement of structural elements, armor plates, and hardware on the vehicle's under-body may result in different deformations depending on where the mine is detonated. Thus, the team identifies six under-body detonation locations, illustrated in Figure 1 that may provide unique deformations: left/back, left/middle, left/front, right/back, right/middle, and right/front. The program would like to be able to detect a difference in deformation between any two of the six detonation locations. The necessity for making these comparisons was driven by careful consideration of how mines affect vehicles.

Intelligence analysts believe that the vehicle is most likely to encounter two variants of the mine type (variant A and B). The program would like to discover if mine variant significantly impacts deformation. This discovery could be critical towards informing military tactics. Additionally, the engineering team would like to determine whether the effect of mine variant on deformation changes as the detonation location changes.

Test Objectives: Detect a difference in deformation between any two of the six detonation Locations, detect a difference in deformation due to Mine Variant, and detect a change in the effect of Mine Variant as Location changes.

After the test program agrees on the objectives, the test planner sets out to design the experiment. Given the established test objectives, the test planner constructs a twice replicated factorial experiment using a two-level factor for Mine Variant and a six-level Location factor. Based on the program's initial allocation of test resources, which accommodate 24 blast events, the test planner finds that the power associated with testing the significance of the Location factor and Mine Variant by Location interaction is unsatisfactorily low, 38% (see Case study example).

Then, the test planner discovers a cost-cutting measure whereby reparameterizing the six-level factor into two factors, a two-level Side factor and a three-level Position factor as show in Figure 2, substantially increases power for the same number of test runs, $\geq$ 89% (see Case study example).

How could this happen? What information was lost? Is the tester still addressing the original objective?

The reparametrization of the factor space changed the hypothesis tests and thus the test objectives. The main effect hypothesis tests in the second proposal no longer correspond to detecting a difference in any two of the six detonation locations on the vehicle nor does the inclusion of a three-way interaction term in the model correspond to detecting a change in the effect of mine Variant as location changes. The tests on the

Side and Position main effects only allow for the detection of a difference between the left and right side, and a difference between any two of the three Position levels. The interaction between these two factors only allows for detecting the effect of one factor (e.g., Side) differing among the levels of the other factor (e.g., Position).

In this paper, we provide the details to demonstrate that the reparameterization results in a different set of hypothesis tests and different test objectives. To do so, we provide a brief review of hypothesis testing and power calculations. Next, we provide the model form used in this paper and review the unified effect size approach (e.g., Oehlert and Whitcomb (2001)), which is useful for comparing reparameterizations of a factor space for a given experimental design. In general, the unified approach is useful for comparing competing designs that have differing sample sizes and varying degrees of imbalance. Following this discussion, we introduce power calculations for the general linear hypothesis test. Lastly, we return to the case study example and resolve the above questions. We discuss the results of the case study and close with a few concluding remarks.

Model formulation

Power depends on many aspects of a designed experiment, including sample size, effect size, and confidence level. Power also depends on the model formulation (i.e., main effects, interactions, polynomial terms, etc.), which is a central theme of this paper. For well-designed experiments, inferences are often made using analysis of variance, which tests for the significance of model effects. In this paper, we define a model effect to be a coefficient or group of coefficients from the model. Power reflects the probability of concluding that the model effect significantly impacts the response variable, when indeed it does.

The model form considered in this paper focuses on effects that comprise nominal categorical factors and their interactions. The expense, limited sample sizes, and in some cases the inability to precisely control factor settings in operational defense testing often leads to broad characterizations that use categorical factors instead of continuous ones.

The increase in power that arises by reparameterizing the factor space can be understood by carefully considering the model formulation. Specifically, attention must be paid to the interpretation and meaning of a model effect and its associated hypothesis test.

We start our review with the ANOVA model, which is useful in this paper for calculating effect sizes. We then transition to reviewing the ANOVA model in matrix notation, which we call the "regression model." We use the regression model notation to calculate power.

ANOVA model

Consider the two-way ANOVA model in Eq. [1]. The model includes two factors (ρ_i and τ_j) with a and b levels, respectively, and the interaction, $\rho\tau_{ij}$. The overall mean is denoted by μ_0.

$$E(y_{ij}) = \mu_{ij} = \mu_0 + \rho_i + \tau_j + \rho\tau_{ij} \ , \quad i = 1, ..., a, \quad j = 1, ...b \tag{1}$$

One aspect of the ANOVA model is that it is overparameterized, which implies that the estimates of the parameters are not unique. To resolve this ambiguity, constraints or side conditions are imposed on the parameters. A commonly used set of conditions are the sum-to-zero constraints, shown below:

$$\sum_{i=1}^{a} \rho_i = 0 \ , \tag{2}$$

$$\sum_{j=1}^{b} \tau_j = 0 \ , \tag{3}$$

$$\sum_{j=1}^{b} (\rho\tau)_{ij} = 0 \quad i = 1, ..., a \ , \tag{4}$$

and

$$\sum_{i=1}^{a} (\rho\tau)_{ij} = 0 \quad j = 1, ..., b \ . \tag{5}$$

Regression model

The constraints in Eqs. [2] through [5] imply that an a-level main effect is sufficiently described by a – 1 regression model coefficients. For the a-level main effect parameter ρ_i, the corresponding coefficient vector β_ρ is of size $(a-1) \times 1$. Let ρ represent the vector of the parameter ρ_i as $i = 1, 2..., a$. Then, using the notation of Hocking (2013), the relationship between the ANOVA model parameter vector ρ and the coefficient vector β_ρ is

$$\rho = \Delta_a^T \beta_\rho \ , \tag{6}$$

where, the contrast matrix is $\Delta_a = (I_{a-1} | -J_{a-1})$, and I_{a-1} is the identity matrix of size $(a-1) \times (a-1)$, and

J_{a-1} is a vector of ones of size $(a-1)\times 1$. For example, if $a=3$, then Eq. [6] is

$$\begin{bmatrix}\rho_1\\ \rho_2\\ \rho_3\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & 1\\ -1 & -1\end{bmatrix}\begin{bmatrix}\beta_{\rho 1}\\ \beta_{\rho 2}\end{bmatrix}. \quad [7]$$

Using this notation, we can express the ANOVA model in Eq. [1] in matrix notation as

$$E(y)=\mu=X\beta\ , \quad [8]$$

where $y=\mu+\epsilon$ and $\epsilon\sim N(0,\sigma^2 I)$. The model matrix $\boldsymbol{X}$ can be partitioned into the following sub-matrices according to each model effect.

$$X=\left(J_a\otimes J_b \ \mid\ \Delta_a^T\otimes J_b \ \mid\ J_a\otimes\Delta_b^T \ \mid\ \Delta_a^T\otimes\Delta_b^T\right) \quad [9]$$

where $J_a\otimes J_b$ corresponds to the intercept, $\Delta_a^T\otimes J_b$ corresponds to the main effect parameter ρ_i, $J_a\otimes\Delta_b^T$ corresponds to the main effect parameter τ_j, and $\Delta_a^T\otimes\Delta_b^T$ corresponds to the interaction parameter $\rho\tau_{ij}$.[1]

For example, let $a=3$ and $b=2$. The model matrix is

$$X=\begin{bmatrix}1 & 1 & 0 & 1 & 1 & 0\\ 1 & 1 & 0 & -1 & -1 & 0\\ 1 & 0 & 1 & 1 & 0 & 1\\ 1 & 0 & 1 & -1 & 0 & -1\\ 1 & -1 & -1 & 1 & -1 & -1\\ 1 & -1 & -1 & -1 & 1 & 1\end{bmatrix}, \quad [10]$$

where $J_a\otimes J_b$ is equal to the first column of $\boldsymbol{X}$, $\Delta_a^T\otimes J_b$ is equal to the second and third columns, $J_a\otimes\Delta_b^T$ is equal to the fourth column, and $\Delta_a^T\otimes\Delta_b^T$ is equal to the last two columns of $\boldsymbol{X}$. Focusing on $\Delta_a^T\otimes J_b$, for example, this can be written out as

$$\begin{aligned}\Delta_a^T\otimes J_b&=\begin{bmatrix}1 & 0\\ 0 & 1\\ -1 & -1\end{bmatrix}\otimes\begin{bmatrix}1\\ 1\end{bmatrix}\\ &=\begin{bmatrix}1\begin{bmatrix}1\\ 1\end{bmatrix} & 0\begin{bmatrix}1\\ 1\end{bmatrix}\\ 0\begin{bmatrix}1\\ 1\end{bmatrix} & 1\begin{bmatrix}1\\ 1\end{bmatrix}\\ -1\begin{bmatrix}1\\ 1\end{bmatrix} & -1\begin{bmatrix}1\\ 1\end{bmatrix}\end{bmatrix}=\begin{bmatrix}1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\\ -1 & -1\\ -1 & -1\end{bmatrix}.\end{aligned} \quad [11]$$

[1]The symbol $\otimes$ denotes the Kronecker product. If A is an $m\times n$ matrix and $\boldsymbol{B}$ is an $p\times q$ matrix, then the Kronecker product works by taking every element in $\boldsymbol{A}$ multiplied by the entire matrix $\boldsymbol{B}$, such that $A\otimes B$ is an $mp\times nq$ block matrix.

Similar to how the model matrix is partitioned, we can also partition the coefficient vector β into sub-vectors for each model effect, expressed as

$$\beta^T=\begin{bmatrix}\beta_0 & \beta_\rho^T & \beta_\tau^T & \beta_{\rho\tau}^T\end{bmatrix}, \quad [12]$$

where β_0 is the intercept coefficient.

Continuing with the example where $a=3$ and $b=2$, Eq. [13] shows the relationship between the ANOVA model and regression model in matrix notation.

$$\begin{aligned}\begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \mu_{21}\\ \mu_{22}\\ \mu_{31}\\ \mu_{32}\end{bmatrix}&=\begin{bmatrix}\mu_0+\rho_1+\tau_1+\rho\tau_{11}\\ \mu_0+\rho_1+\tau_2+\rho\tau_{12}\\ \mu_0+\rho_2+\tau_1+\rho\tau_{21}\\ \mu_0+\rho_2+\tau_2+\rho\tau_{22}\\ \mu_0+\rho_3+\tau_1+\rho\tau_{31}\\ \mu_0+\rho_3+\tau_2+\rho\tau_{32}\end{bmatrix}\\ &=\begin{bmatrix}1 & 1 & 0 & 1 & 1 & 0\\ 1 & 1 & 0 & -1 & -1 & 0\\ 1 & 0 & 1 & 1 & 0 & 1\\ 1 & 0 & 1 & -1 & 0 & -1\\ 1 & -1 & -1 & 1 & -1 & -1\\ 1 & -1 & -1 & -1 & 1 & 1\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_{\rho 1}\\ \beta_{\rho 2}\\ \beta_{\tau 1}\\ \beta_{\rho\tau 1}\\ \beta_{\rho\tau 2}\end{bmatrix}\end{aligned} \quad [13]$$

Unified effect size definition

The effect size plays a crucial role in a power calculation. Typically, the null hypothesis states the model effect is equal to zero, while the alternative hypothesis states it is non-zero. Borrowing nomenclature from the previous section, a hypothesis test on a main effect, ρ_i, is

$$H_0:\begin{bmatrix}\beta_{\rho 1} & \beta_{\rho 2}\end{bmatrix}=\begin{bmatrix}0 & 0\end{bmatrix}\ H_1:\begin{bmatrix}\beta_{\rho 1} & \beta_{\rho 2}\end{bmatrix}\neq\begin{bmatrix}0 & 0\end{bmatrix} \quad [14]$$

which can be written more succinctly as

$$H_0:\beta_\rho=0 \qquad H_1:\beta_\rho\neq 0\ . \quad [15]$$

The effect size is commonly known as the shift in the model effect between the null and alternative hypothesis. Since the null hypothesis typically assumes the model effect is equal to zero under the null, the effect size is simply the value of the model effect under the alternative hypothesis. A challenge with defining an effect size is that there are numerous ways one could accomplish this task, which could all lead to different power results. To provide a unique power estimate for a given experimental design, one needs a unique effect size definition.

Table 1. The candidate search set for Location in the first design proposal.

Config.	1	2	3	4	5	6	7	⋯	13	14	15
τ_1	$\delta/2$	$\delta/2$	$\delta/2$	$\delta/2$	$\delta/2$	0	0		0	0	0
τ_2	$-\delta/2$	0	0	0	0	$\delta/2$	$\delta/2$		0	0	0
τ_3	0	$-\delta/2$	0	0	0	$-\delta/2$	0		0	0	0
τ_4	0	0	$-\delta/2$	0	0	0	$-\delta/2$		$\delta/2$	$\delta/2$	0
τ_5	0	0	0	$-\delta/2$	0	0	0		$-\delta/2$	0	$\delta/2$
τ_6	0	0	0	0	$-\delta/2$	0	0		0	$-\delta/2$	$-\delta/2$

Table 2. The candidate search set for the Mine Variant by Location interaction in the first design proposal.

Config.	1	2	3	4	5	6	7	⋯	13	14	15
$\rho\tau_{11}$	$\delta/2$	$\delta/2$	$\delta/2$	$\delta/2$	$\delta/2$	0	0		0	0	0
$\rho\tau_{12}$	$-\delta/2$	0	0	0	0	$\delta/2$	$\delta/2$		0	0	0
$\rho\tau_{13}$	0	$-\delta/2$	0	0	0	$-\delta/2$	0		0	0	0
$\rho\tau_{14}$	0	0	$-\delta/2$	0	0	0	$-\delta/2$		$\delta/2$	$\delta/2$	0
$\rho\tau_{15}$	0	0	0	$-\delta/2$	0	0	0		$-\delta/2$	0	$\delta/2$
$\rho\tau_{16}$	0	0	0	0	$-\delta/2$	0	0		0	$-\delta/2$	$-\delta/2$
$\rho\tau_{21}$	$-\delta/2$	$-\delta/2$	$-\delta/2$	$-\delta/2$	$-\delta/2$	0	0		0	0	0
$\rho\tau_{22}$	$\delta/2$	0	0	0	0	$-\delta/2$	$-\delta/2$		0	0	0
$\rho\tau_{23}$	0	$\delta/2$	0	0	0	$\delta/2$	0		0	0	0
$\rho\tau_{24}$	0	0	$\delta/2$	0	0	0	$\delta/2$		$-\delta/2$	$-\delta/2$	0
$\rho\tau_{25}$	0	0	0	$\delta/2$	0	0	0		$\delta/2$	0	$-\delta/2$
$\rho\tau_{26}$	0	0	0	0	$\delta/2$	0	0		0	$\delta/2$	$\delta/2$

The unified effect size approach (e.g., Oehlert and Whitcomb (2001)) provides the necessary constraints to obtain a unique effect size definition. The unified approach sizes the effect in terms of the parameter in the ANOVA model and then solves for the coefficients in the regression model. The unified effect size approach contains two parts, with each part placing a constraint on the effect size definition.

Unified Approach Part 1. Define the effect size as the range of the mean response across the design space due to that effect.

As an example of part one, consider a test on a main effect, ρ. If we define ρ under H_1 as $\rho = [-\delta/2 \quad \delta/2 \quad 0]$, then the effect size is δ. Similarly, if we define ρ under H_1 as $\rho = [-\delta/2 \quad \delta/2 \quad \delta/2]$, then the effect size is also δ. Following this trend, there are an infinite number of ways to define ρ to achieve an effect size of δ. Thus, the first part of the unified approach alone is unsatisfactory for obtaining a unique estimate of power.

Part two provides a further constraint by reporting the minimum estimate of power. It requires a search among all configurations of ρ that have a range of δ to find and report the minimum power.

Unified Approach Part 2. Report as power for a given size effect the smallest possible power among all those effects with the given size.

For computationally efficiency, to find the minimum power, the search is constrained to a candidate set of effect sizes. When sizing a test, Dean and Voss (1999, p. 52) note, "the hardest situation to detect is that in which the effects of two of the factor levels … differ by [a certain amount], and the others are all equal and midway between." By the sum-to-zero constraint for main effects (i.e., Eq. [2] or [3]), to achieve an effect size equal to δ implies that the effect of one of the levels should equal $\delta/2$, another should equal $-\delta/2$, and the effect of all other levels in that factor should equal zero. A set that includes an individual effect size for each unique pair of levels is thus the practical candidate set to use to search for minimum power.

An example of a candidate search set for a main effect can be constructed as follows. For this example, let τ represent the six-level location factor that the test planner originally conceived. In general, for a b-level main effect, one can configure "b choose 2" or $\binom{b}{2}$ effect sizes in the candidate set. For a six-level factor there are $\binom{6}{2} = 15$ effect size configurations. The candidate search set is shown in Table 1.

A similar approach can be taken for searching the candidate set of effect sizes for two-factor interactions. Similar to before, to obtain a candidate set that yields low power one must set to zero the effect of as many components of the two-factor interaction as possible. According to the sum-to-zero constraints for two-factor interactions (i.e., Eq. [4] or [5]), to obtain a non-zero effect size, at minimum, four components of the interaction ANOVA parameter must be non-zero. Thus, extending the argument of Dean and Voss to two-factor interactions, set the effects of four components to some value that is non-zero and set the

Table 3. The test design factors, levels, and ANOVA model parameters, shown in parentheses, for the two proposals.

	Location Proposal		Side x Position Proposal		
Factors	Mine	Location	Mine	Side	Position
Levels	A (m_1) B (m_2)	Left/Back (l_1) Left/Middle (l_2) Left/Front (l_3) Right/Back (l_4) Right/Middle (l_5) Right/Front (l_6)	A (m_1) B (m_2)	Left (s_1) Right (s_2)	Back (p_1) Middle (p_2) Front (p_3)

Table 4. The ANOVA models.

Location Proposal	$\mu_{ij} = \mu_0 + m_i + l_j + ml_{ij}$ $i = 1, 2\ ,\quad j = 1, \ldots, 6$
Size × Position Proposal	$\mu_{ijk} = \mu_0 + m_i + s_j + p_k + ms_{ij} + mp_{ik} + sp_{jk} + msp_{ijk}$ $i = 1, 2\ ,\quad j = 1, 2\ ,\quad k = 1, 2, 3$

Table 5. Regression Model Coefficients.

Location Proposal	$\beta = [\,\beta_0 \quad \beta_m \quad \beta_l^T \quad \beta_{ml}^T\,]^T$
Size × Position Proposal	$\beta = [\,\beta_0 \quad \beta_m \quad \beta_s \quad \beta_p^T \quad \beta_{ms} \quad \beta_{mp}^T \quad \beta_{sp}^T \quad \beta_{msp}^T\,]^T$

Table 6. Model matrices.

Location	$\begin{bmatrix} A_l = (& J_2 \otimes J_6 & \mid & \Delta_2^T \otimes J_6 & \mid \\ & J_2 \otimes \Delta_6^T & \mid & \Delta_2^T \otimes \Delta_6^T &) \end{bmatrix}$
Proposal	$\begin{bmatrix} A_{sp} = (& J_2 \otimes J_2 \otimes J_3 & \mid & \Delta_2^T \otimes J_2 \otimes J_3 & \mid \\ & J_2 \otimes \Delta_2^T \otimes J_3 & \mid & J_2 \otimes J_2 \otimes \Delta_3^T & \mid \\ & \Delta_2^T \otimes \Delta_2^T \otimes J_3 & \mid & \Delta_2^T \otimes J_2 \otimes \Delta_3^T & \mid \\ & J_2 \otimes \Delta_2^T \otimes \Delta_3^T & \mid & \Delta_2^T \otimes \Delta_2^T \otimes \Delta_3^T &) \end{bmatrix}$
Size × Position Proposal	

effects of all other components in the interaction equal to zero to obtain minimum power. Moreover, to achieve an effect size of δ, and given that only four component effects in the interaction are non-zero, the sum-to-zero constraints require that two component effects are set at $\delta/2$ and the other other two are set at $-\delta/2$. A set of effect sizes for the two-factor interaction that includes all unique quartets is a practical candidate set to use to search for minimum power.

In general, using this approach to construct effect sizes for an interaction between an a-level and b-level factor, one can construct $\binom{a}{2} \times \binom{b}{2}$ different effect sizes. Returning to the test planner's first proposal, for the interaction between the two-level and six-level factors, leads to $\binom{2}{2} \times \binom{6}{2} = 15$ effect sizes in the candidate search set, as shown in Table 2. For further details on the search method, we refer the reader to Oehlert and Whitcomb (2001).

Power for the general linear hypothesis test

The general linear hypothesis test provides a flexible way to construct a single F-test or simultaneous F-tests on one or more model effects (Rencher and Schaalje 2008). The null and alternative of the general linear hypothesis test are

$$H_0 : C\beta = t \qquad H_1 : C\beta \neq t \qquad [16]$$

where β is the $(k+1) \times 1$ vector of coefficients, and k is the number of coefficients in the model excluding the intercept. The $\boldsymbol{t}$ vector specifies the hypothesized constant value of the effect tested and has size $q \times 1$; note, in practice, $\boldsymbol{t}$ is almost always set to $\mathbf{0}$. The $\boldsymbol{C}$ matrix isolates or constrains the coefficients or combination of coefficients tested and has size $q \times (k+1)$, where $q \leq k+1$. In other words, q is the number of simultaneous hypotheses being tested. For example, if $H_0 : \beta_1 = 0$ then $q = 1$; if $H_0 : \beta_1 = \beta_2 = 0$ then $q = 2$; if $H_0 : \beta_1 - 2\beta_3 = \beta_2 - \beta_3 = -\beta_4 = 0$ then $q = 3$.

When performing the hypothesis test on the coefficients, we are formally testing if the model coefficients β are significantly different from some constant, $\boldsymbol{t}$. The power of the test is the probability that the test correctly rejects the null hypothesis, when the alternative hypothesis is true. The test statistic for analysis of variance is an F-statistic which depends on the collected data $\boldsymbol{y}$, the model matrix $\boldsymbol{X}$, and the coefficients estimated from the collected data $\hat{\beta} = (X'X)^{-1}X'y$, and is calculated as

$$F = \frac{\left(C\hat{\beta} - t\right)' \left[C(X\prime X)^{-1}C\prime\right]^{-1} \left(C\hat{\beta} - t\right)/q}{y' \left[I - X(X\prime X)^{-1}X'\right] y/(n-k-1)} \ . \qquad [17]$$

If the null hypothesis is true, then F follows a central F distribution with q numerator degrees of freedom, and $n-k-1$ denominator degrees of freedom, where n is the sample size and k is the number of coefficients in the model (excluding the intercept coefficient).

If the null hypothesis is false, then F follows a non-central F distribution with non-centrality parameter λ, and q numerator and $n-k-1$ denominator degrees of freedom. The non-centrality parameter is

$$\lambda = \left(C\hat{\beta} - t\right)' \left[C(X\prime X)^{-1}C\prime\right]^{-1} \left(C\hat{\beta} - t\right)/\sigma^2 \ . \qquad [18]$$

Note, for the purpose of sizing a test, we define an effect size by assigning a value for the coefficients in Eq. [18].

The power of the test is equal to

$$P\left(F \geq \hat{F}_\alpha\right) = 1 - \tilde{F}\left(\hat{F}_{\alpha,q,n-k-1}, q, n-k-1, \lambda\right) \ , \qquad [19]$$

where $\hat{F}$ is the F central quantile function which provides the critical F value evaluated at the $(1-\alpha)$th

Table 7. Hypothesis tests for each proposal.

Location Proposal		Size × Position Proposal	
Hypothesis	C	Hypothesis	C
$H_0: \beta_m = 0$	$[0 \quad 1 \quad 0_{1\times10}]$	$H_0: \beta_m = 0$	$[0 \quad 1 \quad 0_{1\times10}]$
$H_0: \beta_l = 0$	$[0_{5\times2} \quad I_5 \quad 0_{5\times5}]$	$H_0: \beta_s = 0$	$[0_{1\times2} \quad 1 \quad 0_{1\times9}]$
$H_0: \beta_{ml} = 0$	$[0_{5\times7} \quad I_5]$	$H_0: \beta_p = 0$	$[0_{2\times3} \quad I_2 \quad 0_{2\times7}]$
		$H_0: \beta_{ms} = 0$	$[0_{1\times5} \quad 1 \quad 0_{1\times6}]$
		$H_0: \beta_{mp} = 0$	$[0_{2\times6} \quad I_2 \quad 0_{2\times4}]$
		$H_0: \beta_{sp} = 0$	$[0_{2\times8} \quad I_2 \quad 0_{2\times2}]$
		$H_0: \beta_{msp} = 0$	$[0_{2\times10} \quad I_2]$

quantile, and $\tilde{F}$ is the non-central F distribution function evaluated at the critical F value.

Case study example

Having summarized the analytical foundations, we return to our case study example. In this section, we highlight a Type III error that test planners might commit in an attempt to reduce test size and resources. Our case study investigates the two test design proposals.

In each proposal, the test planner chooses static deformation as the response variable of interest, which is normally distributed and quantifies the deformation of the vehicle's armor due to the mine blast. Each experimental run consists of a single detonation event, resulting in one measurement of deformation, measured in inches. We refer to the first proposal as the "Location Proposal", which differentiates between the six detonation locations using a single factor. We refer to the second proposal as the "Side × Position Proposal," which reparameterizes the six-level location factor into two separate factors. The two proposals include the same Mine Variant factor. Table 3 lists the factors and levels for each proposal.

The test planner chooses a main effects plus two-factor interaction model for the Location Proposal, and includes an additional three-factor interaction in the Side by Position proposal. The ANOVA models for each proposal appear in Table 4.

The effect sizes are first defined in terms of the parameters of the ANOVA model and are then converted into regression model coefficients. In each proposal the coefficient vector β has size 12×1. Each coefficient vector can be partitioned by model effects comprising main effects and two-way interactions according to Table 5.

The design for each proposal is the same, which is a duplicated full factorial experiment resulting in 24 runs. Let A_l and A_{sp} denote the single replicate full factorial model matrix for the Location Proposal and Side by Position Proposal, respectively, each of size 12×12. The model matrices are $X_l = [A_l^T \mid A_l^T]^T$ and $X_{sp} = [A_{sp}^T \mid A_{sp}^T]^T$, where the single replicate full factorial model matrices are constructed according to Table 6.

Table 8. Coefficients assuming H_1 is true.

Location Proposal	Size × Position Proposal
$\beta_m = 1/2$	$\beta_m = 1/2$
$\beta_l = [1/2 \quad 0 \quad 0 \quad 0 \quad 0]^T$	$\beta_s = 1/2$
$\beta_{ml} = [1/2 \quad 0 \quad -1/2 \quad 0 \quad 0]^T$	$\beta_p = [1/2 \quad -1/2]^T$
	$\beta_{ms} = 1/2$
	$\beta_{mp} = [1/2 \quad -1/2]^T$
	$\beta_{sp} = [1/2 \quad -1/2]^T$
	$\beta_{msp} = [1/2 \quad -1/2]^T$

After defining the models, the test planner is ready to calculate the power associated with assessing the test objectives. Recall, the test objectives are to detect a difference in deformation between any two of the six detonation Locations, detect a difference in deformation due to Mine variant, and detect a change in the effect of Mine Variant as Location changes. Following typical procedures using statistical software, the test planner assumes that, by calculating power for main effects and interactions in each proposal, they are dutifully addressing the test objectives.

The test planner constructs the hypothesis tests using the format of the general linear hypothesis test. Table 7 shows the matrix $\boldsymbol{C}$ associated with each hypothesis test on the main effects and interactions for the two proposals. The hypothesis tests in this table imply that $\boldsymbol{t}$ is equal to zeros in Eq. [18].

Next, the test planner defines the effect sizes. An effect size is defined for each hypothesis test and represents the value of the coefficients tested assuming H_0 is false and H_1 is true. Using the unified approach, the test planner first defines the effect size in terms of parameters of the ANOVA model and then converts those parameters into coefficients for the regression model.

The test planner selects an effect size of 1 inch. In the Location Proposal, letting $\boldsymbol{m}$, $\boldsymbol{l}$, and $\boldsymbol{ml}$ be the vectors of the ANOVA model parameters for m_i, l_j, and ml_{ij} the effect size definition implies the range of $\boldsymbol{m}$, $\boldsymbol{l}$, or $\boldsymbol{ml}$ is equal to 1 inch. Using a similar approach in the Side by Position Proposal, the effect size implies the range of $\boldsymbol{m}$, $\boldsymbol{s}$, $\boldsymbol{p}$, $\boldsymbol{ms}$, $\boldsymbol{mp}$, $\boldsymbol{sp}$, or $\boldsymbol{msp}$ is equal to 1 inch.

Part two of the unified approach requires a search among the candidate set of effect sizes for the particular effect size that yields minimum power. The candidate search is unnecessary in this case study because the test designs are completely balanced. For balanced designs, power for each effect size within a candidate set is identical. It is only with unbalanced designs that the candidate search is necessary to provide a unique estimate of power.

Table 9. Non-centrality parameter and power for each hypothesis test.

Location Proposal				Size × Position Proposal			
Hypothesis	q	λ	Power	Hypothesis	q	λ	Power
$H_0 : \beta_m = 0$	1	24	.99	$H_0 : \beta_m = 0$	1	24	.99
$H_0 : \beta_l = 0$	5	8	.38	$H_0 : \beta_s = 0$	1	24	.99
$H_0 : \beta_{ml} = 0$	5	8	.38	$H_0 : \beta_p = 0$	2	16	.89
				$H_0 : \beta_{ms} = 0$	1	24	.99
				$H_0 : \beta_{mp} = 0$	2	16	.89
				$H_0 : \beta_{sp} = 0$	2	16	.89
				$H_0 : \beta_{msp} = 0$	2	16	.89

The test planner defines the individual ANOVA model parameter vectors to satisfy part one of the unified approach. To illustrate for ***l*** in the first proposal, the test planner arbitrarily chooses the fifth configuration (see Table 1) as shown in Eq. [20]. Although the fifth configuration from Table 1 was chosen, any configuration could have been selected; each configuration gives the same power because the design is balanced.

$$l = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & -1/2 \end{bmatrix}^T \quad [20]$$

The equation that converts the ANOVA model parameter to the regression model coefficients is,

$$\beta_l = \left(\Delta_6^T\right)^{-1} l = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 \end{bmatrix}^T . \quad [21]$$

The conversion for two- and three-factor interactions follows a similar process. Take for example the ***ml*** interaction in the first proposal. The test planner arbitrarily chooses the second configuration from Table 2 so that the ANOVA model parameter vector is

$$\mathrm{ml} = \begin{bmatrix} 1/2 & 0 & -1/2 & 0 & 0 & 0 & -1/2 & 0 & 1/2 & 0 & 0 & 0 \end{bmatrix}^T . \quad [22]$$

The equation that converts the ANOVA model parameter to regression model coefficients is

$$\beta_{ml} = \left(\Delta_2^T \otimes \Delta_6^T\right)^{-1} \mathrm{ml} = \begin{bmatrix} 1/2 & 0 & -1/2 & 0 & 0 \end{bmatrix}^T . \quad [23]$$

The test planner repeats this calculation process for each hypothesis test. The resulting coefficients for each proposal appear in Table 8.

Having defined the effect sizes, the test planner calculates the power. The assumed confidence level is 95% ($\alpha = 0.05$). Next, the test planner calculates the non-centrality parameter λ using Eq. [18]. In this equation, σ is the root mean-squared error, representing the overall "noise" in the experiment. Based on observations from previous testing that had been executed under similar conditions, the test planner assumes σ is equal to 0.5 inches, which implies a "signal-to-noise" ratio equal to 2 (recall the effect size or "signal" is 1 inch). Finally, the test planner calculates the power using Eq. [19]. The numerator degrees of freedom, non-centrality parameter, and the power for each hypothesis test and proposal are shown in Table 9.

After completing the power calculations for both proposals, the test planner prepares briefing slides and presents the results to a room of non-statistically oriented engineers, managers, and military personnel. The premise of the results, although omitted from the presentation, is that both proposals address the test objectives. Without any discussion about the connection between the hypothesis tests and test objectives, the test planner quickly arrives at the power results for each proposal. The choice of proposal becomes clear. For the same number of runs, the Side by Position proposal provides no less than 89% power for all main effects and interactions, compared to the 38% power for the Location proposal. Impressed by the savings in test resources, the test program agrees to the Side by Position Proposal and commits a Type III error.

Discussion

Given the completion of the case study example, we discuss what went wrong and how one might diagnose the problem. The mistake was reparameterizing the factor space. The main effects and interaction hypothesis tests in the Location proposal correctly addressed the test objectives; however, the reparameterization changed the hypothesis tests.

The main effect and interaction hypothesis tests in the Side by Position proposal do not address the test objectives. In truth, the tests on the Side and Position main effects allow for the detection of a difference between the left and right side, and a difference between any two of the three Position levels, respectively. Neither of these hypothesis tests, nor the interaction test, allows for the detection of a difference between any two of the six locations on the vehicle.

Reparameterizing the factor space is not wrong in itself, as long as careful consideration is given to the hypothesis tests. For instance, the test planner could have conducted a hypothesis test in the Side by

Position proposal that is equivalent to the test on the Location main effect in the Location proposal. Recall the hypothesis test $\beta_l = 0$ from before, where Table 9 shows that $q = 5$, $\lambda = 8$, and power equals 38%. Finding the equivalent hypothesis test for the Side by Position proposal, as we will now show, clarifies the similarity between models and demonstrates that there is no artificial gain in power using the Side by Position proposal.

First, we must construct an equivalent effect size between proposals. Start by calculating the change in the mean response due to the location main effect in the Location proposal. That is,

$$\mu = A_l \begin{bmatrix} \beta_0 & \beta_m & \beta_l^T & \beta_{ml}^T \end{bmatrix}^T. \quad [24]$$

In Eq. [24], A_l is from Table 6, the values of β_l^T are found in Table 8, and since we would like to isolate the location effect in this example we set β_0, β_m, and β_{ml}^T equal to zeros.

The equivalent effect size in terms of the coefficients for the Side by Position proposal is found as

$$\beta = A_{sp}^{-1}\mu \ , \quad [25]$$

where A_{sp} is from Table 6. Inserting μ from Eq. [24] into Eq. [25], results in

$$\begin{aligned} \beta &= \begin{bmatrix} \beta_0 & \beta_m & \beta_s & \beta_p^T & \beta_{ms} & \beta_{mp}^T & \beta_{sp}^T & \beta_{msp}^T \end{bmatrix}^T \\ \beta &= \begin{bmatrix} 0 & 0 & 1/6 & 1/4 & 0 & 0 & 0 & 0 & 1/12 & -1/6 & 0 & 0 \end{bmatrix}^T , \end{aligned} \quad [26]$$

which is the effect size of the location main effect from the Location proposal expressed in terms of coefficients for the Side by Position proposal. The non-zero components of Eq. [26] correspond to β_s, β_p^T, and β_{sp}, which collectively have the same number of degrees of freedom as the location factor in the Location proposal (equal to five). Thus, the equivalent five degree of freedom hypothesis test is

$$H_0 : \beta_s = \beta_p = \beta_{sp} = 0 \ , \quad [27]$$

which has matrix $\boldsymbol{C}$ that takes the form

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}. \quad [28]$$

The power calculation is then carried out as before. For $\sigma = 0.5$ and $\alpha = 0.05$, using Eqs. [18] and [19], the non-centrality parameter is 8 and the power of this test is 38%.

Note, the power is identical to the test of the Location main effect in the Location proposal. This result demonstrates that, given an equivalent effect size, each proposal has the same power to assess the hypotheses that addresses the test objectives.

Conclusion

The test planner mistakenly believed that the main effects and interaction hypothesis tests in the Side by Position proposal addressed the test objectives. The perceived high power associated with this proposal led the team to believe the test was adequately resourced. In this particular case, the high power was not used to argue for a smaller test design, but this could and does happen in practice. The proper decision would have been to select the Location proposal, and recognize that the only solution to adequately assessing the test objectives are to add more experimental runs.

The case study highlights the negative consequences of redefining a factor space, but this approach should not be completely discredited. If the test planner and the test program had renegotiated the objectives, such that the objectives aligned with the hypothesis tests in the Side by Position proposal, this could have been a shrewd cost-savings strategy. It is the lack of careful planning that leads to a Type III error.

As a parting thought, it was the unified effect size approach, coupled with the ability to convert effect sizes between model formulations that facilitated the diagnosis of the Type III error. These tools were useful for comparing reparameterizations of a factor space for a fixed experimental design, but they can also be useful for comparing competing designs that have different sample sizes or different degrees of imbalance. The consistent estimate of power from the unified approach enables such comparisons.

About the authors

Thomas H. Johnson is a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses.

Rebecca M. Medlin is a Research Staff Member in the Operational Evaluation Division at the Institute for Defense Analyses.

Laura J. Freeman is an Assistant Director in the Operational Evaluation Division at the Institute for Defense Analyses.

James R. Simpson is Principal of JK Analytics and Consultant at the Institute for Defense Analyses.

References

Dean, A. M., and D. Voss. 1999. *Design and analysis of experiments*. Volume 1. New York, NY: Springer.

Freeman, L. J., A. G. Ryan, J. L. Kensler, R. M. Dickinson, and G. G. Vining. 2013. A tutorial on the planning of experiments. *Quality Engineering* 25 (4):315–32.

Hahn, G. J., W. Meeker, and P. I. Feder. 1976. The evaluation and comparison of experimental designs for fitting regression relationships. *Journal of Quality Technology* 8 (3):140–57.

Hocking, R. R. 2013. *Methods and applications of linear models: Regression and the analysis of variance*. New York, NY: John Wiley & Sons.

Johnson, R. T., G. T. Hutto, J. R. Simpson, and D. C. Montgomery. 2012. Designed experiments for the defense community. *Quality Engineering* 24 (1): 60–79.

Kimball, A. 1957. Errors of the third kind in statistical consulting. *Journal of the American Statistical Association* 52 (278):133–42.

Oehlert, G. W., and P. Whitcomb. 2001. Sizing fixed effects for computing power in experimental designs. *Quality and Reliability Engineering International* 17 (4): 291–306.

Rencher, A. C., and G. B. Schaalje. 2008. *Linear models in statistics*. New York, NY: John Wiley & Sons.