Employing Power to 'Right-Size' Design of Experiments

Mark J. Anderson, *PE, CQE, MBA*
Principal and General Manager, Stat-Ease, Inc. Minneapolis, MN

Patrick J. Whitcomb, *PE, MS*
Founding Principal and President, Stat-Ease, Inc. Minneapolis, MN

This article provides insights on how many runs are required to make it very likely that a test will reveal any important effects. Due to the mathematical complexities of multifactor Design of Experiments (DOE) matrices, the calculations for adequate power and precision (Oehlert and Whitcomb 2002) are not practical to do by 'hand' so the focus will be kept at a high level—scoping out the forest rather than detailing all the trees. By example, readers will learn the price that must be paid for an adequately-sized experiment and the penalty incurred by conveniently grouping hard-to-change factors.

Introduction

Let's start with two telling observations provided in an inspiring MOA on DOE for test and evaluation (T&E)—"Using Design of Experiments for Operational Test and Evaluation," Memorandum of Agreement, Nov 24, 2009, Operational Test Agency Technical Directors and the Science Advisor for Operational Test and Evaluation:

- *"DOE provides the scientific and statistical methods needed to rigorously plan and execute tests and evaluate their results."*
- *"A DOE-based test approach will not necessarily reduce the scope of resources for adequate testing."*

Of course, with every good thing (first statement) comes a catch (second one): To achieve the benefits of DOE, the plan must provide enough runs to make detection of problematic effects very likely. Based on anecdotal evidence, such as the story of an unnamed aircraft tester who flew two dozen flights by the seat of his pants and then asked a statistician to develop a DOE on three that remained (mission impossible!), expectations need to be managed on what really must be done for a rigorous experiment.

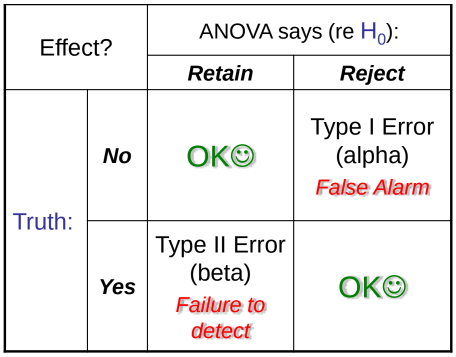
A Primer on Power

Power is the probability of revealing an active effect of size delta (Δ) relative to the noise (σ) as measured by signal to noise ratio (Δ / σ). It should be high—at least 80 %, but beyond 95 % power – or the returns of more runs diminish greatly. In other words, the right size for an experiment design is one with enough runs to fall within the range of 80-95% power for all critical responses.

The flip side of power is the risk of not seeing an important effect—one that exceeds the difference delta. It is a failure in detection. Statisticians refer to this as a Type II error and symbolize it beta (β). Power equals one minus beta ($1-\beta$) times 100.

Another error, which tends to get a lot more attention is the Type I, which is the risk of saying an effect is active when it is really not, that is, a false positive. This is represented as alpha (α). Type I error is assessed by the p-value from analysis of variance (ANOVA), which tests the null hypothesis (H_0) that nothing is significant.

The 'truth table' in Figure 1 lays out four possibilities, two of which are OK—the others being errors either of Type I or Type II.



Effect?		ANOVA says (re H_0):	
		Retain	***Reject***
Truth:	***No***	OK☺	Type I Error (alpha) *False Alarm*
	Yes	Type II Error (beta) *Failure to detect*	OK☺

Figure 1: Truth test illustrating two types of error

A good way to explain these errors, which, as a practical matter, can never be completely eliminated, is to consider the smoke detector in your kitchen. If it goes off when you make perfectly good toast, that is a false alarm—a Type I error. Dialing the sensitivity down (or removing the battery altogether!), to the extent that a real fire fails to set off the alarm, puts you in harm's way by a Type II error. So the take-home message is that you really ought to pay attention to power.

Sizing a Simple Comparative Experiment

The simplest DOE is a comparison of two things, such as one material versus another or two alternate suppliers. Even a statistically-challenged test engineer would naturally set up an experiment with at least four of each in a total of eight runs, thus providing some power from averaging. However, with just the following few bits of intelligence a far more statistical approach can be accomplished for adequately sizing a DOE:

- Specification of the minimal effect of interest, that is, difference in response delta (Δ)—also known as the "signal"
- An estimate of the process variation, that is, standard deviation sigma (σ)—called "noise"
- How much power is needed, that is, the probability of detecting the effect—generally acceptable being set to 80 percent (0.8 on a zero to one scale) or better
- The risk of a false positive outcome, also known as a Type I or error and symbolized by alpha (α)—typically established at 5 percent (0.05)(Kelly 2013).

Entering these inputs into a two-sample power calculator—readily available in statistical software or internet applets such as one posted by a University of Iowa statistics professor (Lenth 2012) —produces an estimate of the required number (n) of runs for each treatment. Preferably the experiment will provide a balance of runs, thus the total number (N) will be double this estimate (2n).

Now let's work through a universal dilemma that illustrates the process of powering up a DOE.

Case Study 1: Experimenting on Which Route to Take

Have you ever wondered if it might be sensible in the long run to take an alternative route to work? Why not try? It may well provide a decisive advantage (Anderson 2013). But, hold on, how many drives will it take for you to see if it really matters one way or the other? First off, consider how much of a difference it would take to get you to budge from your habitual path. Perhaps 2 minutes might be just enough. Let's assume so and shift to plain English with the symbol "d" for this potential difference in means between the routes (current versus alternative).

Next comes the trickier part—guessing at the variation in your drive, that is, the standard deviation. The best way to come up with this number is by simply recording your drive times for 20 or more runs into work. Type these into a calculator or spreadsheet, or enter them into software providing simple-sample data characterization, and compute the "s". For purpose of this exercise, figure this optimistically to be 1 minute, which produces a signal-to-noise ratio of 2. With these parameters, the power for which as Scenario "A" Table 1 shows the number of runs required for adequate power—34 total broken down into 17 each way, which you'd best do in a random order (by the flip a real or virtual coin).

Scenario	Signal (d)	Noise (s)	Signal/Noise	Runs	Power
A	2 min.	1 min.	2.0	12	87.6%
B	2 min.	2 min.	1.0	34	80.7%
C	2 min.	4 min.	0.5	128	80.1%
D	4 min.	4 min.	1.0	34	80.7%

Table 1: Runs required for adequately powering (>80%) a simple-comparative experiment

Scenario B recognizes that the variation of your commuting time might be more than you thought—say twice the standard deviation. Then you'd just have to keep going, but not just twice as many times as one might expect. It turns out that, subject to other statistical factors at the low end, there's an exponential relationship between the signal-to-noise ratio and the number of runs. (Note we are going for a power of at least 80% for our calculations.) You can see this in Scenario C where the noise ratchets up again by another doubling—causing the runs soar to 128. Scenario D lays out a solution that you may not really like, but it may be all that can be done: lower your expectations on what the experiment will uncover by increasing the minimum-detectable signal to 4 minutes. In other words, if the noise does come in at the higher level of 4, then for the 34 runs you might be willing to invest in the design, only an effect of 4 minutes or more is likely to emerge—if it's there at all.

The big advantage of adequately powering an experiment like this is that an insignificant outcome provides a win by eliminating ground from having to be covered again. It may very well turn out that the way you now commute is about as good as or better than the alternative. If so, you came out ahead by eliminating that nagging feeling about the alternate perhaps being better. Often it's this process of elimination, if statistically defensible, that wins the day.

Designing a Two-Level Factorial Experiment with an Adequate Number of Runs

The real advantage for DOE comes with multifactor test plans, particularly the ones where factors are restricted to two levels (Anderson 2007). These designs are most efficient for quantifying main effects via the parallel processing schemes embedded in their test matrices. If done at proper resolution, they also can uncover previously unknown interactions that can be exploited for beneficial effects (synergistic) or studiously avoided when negative (antagonistic). However, the spear will be blunted when too few runs are achieved to reveal meaningful effects. It's vital that power be assessed before going into action with any given test plan. To calculate this requires the same inputs for signal and noise, but, as noted at outset the actual computations can, as a practical matter, only be done by sophisticated software. Let's take a look at an elegant example so you get the big picture.

Case Study 2: Designing a Helicopter that Consistently Flies Long and Lands Accurately with Maximum Precision

Building paper helicopters as an in-class DOE project is well-established for a hands-on learning experience (Box 1992). Figure 2 pictures an example—the top flyer from the trials detailed in this case study.

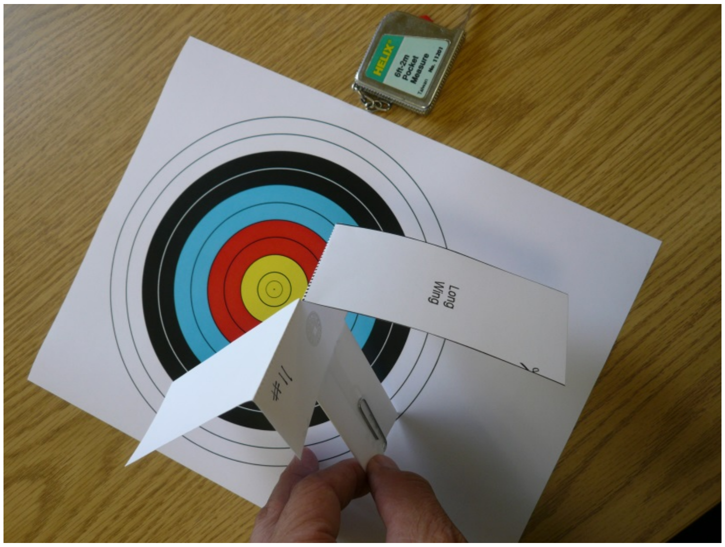


Figure 2: Paper helicopter

Modern DOE programs feature a very tempting option called "split plots" that accommodate hard-to-change factors such as dimensional factors in aircraft construction. These split-plot designs are very tempting, but as you will see in our example, they come with a price in terms of power.

"Oftentimes, complete randomization of all test parameters is extremely inefficient, or even totally impractical. One or more of these parameters (e.g. altitude) can be altered a minimum number of times during test by rearranging the test run order. The proper test strategy in these instances is called a split-plot design."

- Alex Sewell, 53rd Test Management Group, 28th Test and Evaluation Squadron, Eglin AFB (Sewell et al, 2009)

Engineers at Stat-Ease selected the following six factors at the levels shown (details, including templates, available upon request to corresponding author) for a characterization of main effects and possible two-factor interactions in a half-fraction, high-resolution (Res VI) two-level design with 32 runs. The experiment was inspired by news of a supreme paper—Conqueror CX22—made into an airplane that broke the Guinness World Record™ for greatest distance flown (Spangers 2012). This is the high level of the first factor.

- Hard to change (construction)
  a. Paper: 24# Navigator Premium (standard) versus 26.6# Conqueror CX22 (supreme)
  b. Wing Length: Short vs Long
  c. Body Length: Short vs Long
  d. Body Width: Narrow vs Wide
- Easy to change (operation)
  E. Clip: Off vs On
  F. Drop: Bottom vs Top

Notice that, by convention, the hard-to-change (HTC) factors, related to construction of the helicopters, are designated by lower case letters. These factors are grouped in "whole plots"—a term that comes from the field of agronomy where split-plot designs were invented. The other (capitalized) factors can be easily changed and thus randomized in the "sub-plots".

To develop the longest flying, most accurate helicopter the experimenters measured these two responses (Y):

1. Average time in seconds for three drops from ceiling height
2. Average deviation in centimeters from target

The averaging dampened out drop-to-drop variation, which, due to human factors in the release, air drafts and so forth, can be considerable. As shown in Table 2, of the two responses, the distance from target (Y2) produced the lowest signal-to-noise ratio. It was based on a 5 cm minimal deviation of importance relative to a 2 cm standard deviation measured from prior repeatability studies.

Response	Signal	Noise	Signal/Noise
Time avg	0.5 sec	0.15 sec	3.33
Target avg	5.0 cm	2.00 cm	2.50

Table 2: Signal to noise for the two paper helicopter responses

At a 2.5 signal-to-noise ratio a 32-run two-level factorial design generates the power shown in Table 3.

Design	Hard (a-d)	Easy (E, F)
Split plot	82.1%	99.9%
Randomized	97.5%	97.5%

Table 3: Power for main effects for design being done as split plot versus fully randomized

What's really important to see here is that by grouping the HTC factors the experimenters lost power versus a completely-randomized design. (It mattered little in this case, but it must be noted that the other factors gain power.) However, the convenience of only building half the paper helicopters —16 out of the 32 required in the fully-randomized design of experiments— outweighed the loss in power, which in any case remained above the generally-acceptable level of 80%. Thus this test plan, laid out in Table 4, got the 'thumbs-up' from the flight engineers.

Group	Run	a:Paper	b:Wing	c:Body Length	d:Body Width	E:Clip	F:Drop
1	1	Nav Ultra	Long	Short	Narrow	Off	Top
1	2	Nav Ultra	Long	Short	Narrow	On	Bottom
2	3	CX22	Short	Short	Narrow	Off	Top
2	4	CX22	Short	Short	Narrow	On	Bottom
3	5	Nav Ultra	Long	Long	Narrow	Off	Bottom
3	6	Nav Ultra	Long	Long	Narrow	On	Top
…	…	…	…	…	…	…	…
16	31	CX22	Short	Long	Wide	Off	Top
16	32	CX22	Short	Long	Wide	On	Bottom

Table 4: Partial listing (first three groups and the last) of 32-run split-plot test plan

Much more could be said about split plots and the results of this experiment in particular (for what it's worth—CX22 did indeed rule supreme). However, the main point is how power was employed to 'right-size' the test plan while accommodating the need to reduce the builds.

Conclusion

This article makes the case for experimenters being diligent by doing power calculations to assess how many runs are required to make it very likely that their test plan will reveal any important effects. Advancements in software make this easy to do, even for relatively sophisticated designs such as a two-level factorial split plot. For non-orthogonal designs, such as response surface methods (RSM), other methods can be employed for the same purpose, that is, to 'right-size' your experiment design and thus be assured that, whether or not anything emerges significant, the results will be statistically defensible.

Mark J. Anderson, PE, CQE, MBAis a principal and general manager of Stat-Ease, Inc. Prior to joining the firm, he spearheaded an award-winning quality improvement program for an international manufacturer, generating millions of dollars in profit. Mark offers a diverse array of experience in process development, quality assurance, marketing, purchasing, and general management. Mark is also co-author of two books, "DOE Simplified: Practical Tools for Effective Experimentation" and "RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments," and has published numerous articles on design of experiments (DOE). He is also a guest lecturer at the University of MN Chemical Engineering & Materials Science department and the Ohio State University Fisher College of Business. Email: mark@statease.com

Patrick J. Whitcomb, PE, MSis the founding principal and president of Stat-Ease, Inc. Before starting his own business, he worked as a chemical engineer, quality assurance manager, and plant manager. Pat co-authored Design-Ease® software, an easy-to-use program for design of two-level and general factorial experiments and Design-Expert® software, an advanced user's program for response surface, mixture, and combined designs. He's provided consulting on the application of design of experiments (DOE) and other statistical methods for several decades. In addition, Pat is co-author of the books, "DOE Simplified: Practical Tools for Effective Experimentation" and "RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments," and has published many articles on design of experiments (DOE).

References

Anderson, M. J. and P.J. Whitcomb. 2007. *DOE Simplified, Practical Tools for Effective Experimentation, 2nd Edition*. New York, NY. Productivity Press.

Anderson, M. J. 2013. Statistics Point the Way to Save Time Commuting. *Stat-Teaser* September 2013 posted at http://www.statease.com/news/news1309.pdf.

Box, G. E. P. 1992. George's Column: Teaching Engineers Experimental Design with a Paper Helicopter. *Quality Engineering*, 4 (3): 453–459.

Kelly, M. 2013. Emily Dickinson and monkeys on the stair, Or: What is the significance of the 5% significance level. *Significance* 10 (5): 21-22.

Lenth, R. V. 2006-12. Java Applets for Power and Sample Size [Computer software]. Retrieved November, 2013, from http://www.stat.uiowa.edu/~rlenth/Power.

Oehlert, G. W. and P. J. Whitcomb. 2002. Sizing Fixed Effects for Computing Power in Experimental Designs. *Quality and Reliability Engineering International* 17 (4): 291–306.

Sewell, A., Kowalski, M. and J. Simpson. 2009. Standard Operational Test Planning and Analysis Via Nonstandard Designed Experiments. *American Institute of Aeronautics and Astronautics*. 2009-1709 (U.S. Air Force T&E Days 2009).

West, H and J. Spangers. 2012. World record-breaking paper airplane takes flight on Conqueror. Appleton Coated LLC press release posted at http://www.appletoncoated.com/ricofiles/pdf/pressrelease/2012acu_paperairplanes_final.pdf.