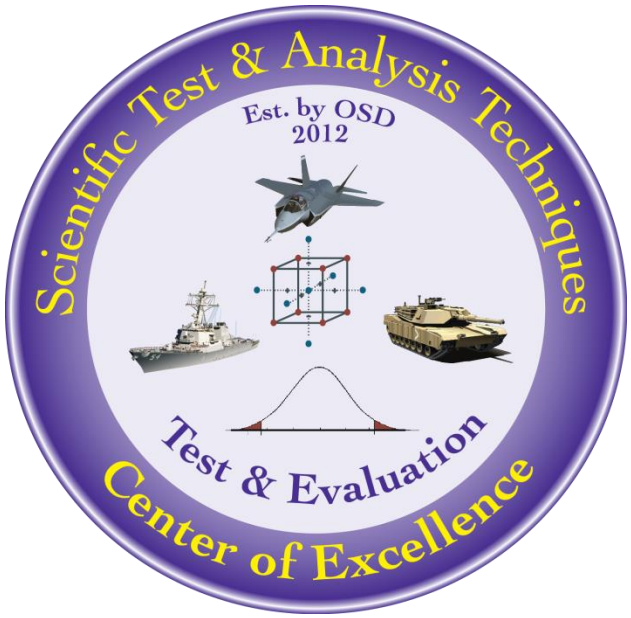
Test Design Comparison and Selection

Authored by: Michael Harman

19 August 2014

Revised 29 Aug 2018



The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Revision 1, 29 Aug 2018, Formatting and minor typographical/grammatical edits.

Executive Summary

The leadership demand for test rigor in Department of Defense (DoD) test and evaluation requires the use of methodical, repeatable, and defensible design selection processes. The use of design of experiments (DOE) increases the rigor in testing, but the DOE planning method does not result in a lone design. In fact, given test objectives, analysis requirements, budget, and other constraints the designer can choose from multiple options to balance the trade space. This paper presents a quantitative and defensible method for comparing and selecting designs among the technical choices available to the practitioner.

Keywords: rigor, design of experiments, design selection

Introduction

Despite the rigor and procedure associated with design of experiments methods, there are still choices to be made in the design selection process. These choices allow the practitioner to address budget, analysis expectations, and constraints and limitations. This paper assumes the reader has a working knowledge of the planning process required to arrive at the type of design necessary to achieve test objectives.

Multiple inputs (beyond factors and levels) are required to define a design, and certain metrics can be useful in evaluating the design utility. Many inputs are set at default values in DOE applications, so it is necessary to consider changing these values to assess their sensitivity on a given design space. Methodically, it can become difficult to assess design differences when more than one input is changed. Using a table like Figure 1 greatly simplifies the process and supports a proper and methodical analysis leading to a final design selection. The practitioner employs DOE software to create and evaluate designs with various inputs and metrics and enters the design output into an Excel spreadsheet column by column.

The row colors are relative based on the values in any particular row so green numbers reflect the more desirable trend. Rows can be formatted red (low) and green (high) [e.g. power, higher being better] or red (high) and green (low) [e.g. number of runs, higher costing more]. This is easily accomplished in Excel using built-in conditional formatting and allows a visual interpretation: “greener” columns indicate a more desirable design. Note that using text in these cells negates the ability to use this coloring function.

Design #	1	2	3	4	5	6
Software Generator	JMP	JMP	JMP	JMP	JMP	JMP
Name	RSM FCD Custom	RSM FCD Custom	RSM FCD Custom	RSM FCD Custom	RSM FCD Custom	RSM FCD Custom
# Factors	4	4	4	4	4	4
Levels	2	2	2	2	2	2
Model Supported	ME, 2FI, Q	ME, 2FI, Q	ME, 2FI, Q	ME, 2FI, Q	ME, 2FI, Q	ME, 2FI, Q
Signal to Noise Ratio	2.0	2.0	2.0	2.0	2.0	2.0
Alpha	0.05	0.05	0.05	0.05	0.05	0.05
Blocks of size	8	8	8	8	8	8
# Center Points	0	4	0	4	6	6
# Repetitions	7	7	14	14	14	24
Total Runs	20	24	30	34	36	56
Power for ME @ S/N	0.63	0.54	0.9	0.92	0.95	0.98
Power for 2FI/Q @ S/N	0.54	0.53	0.8	0.73	0.87	0.94
FDS Pred Err @50%	0.63	0.58	0.45	0.37	0.36	0.24
FDS Pred Err @95%	0.90	0.75	0.58	0.47	0.45	0.30
VIF Avg	4.09	3.73	3.73	3.64	3.45	3.36
VIF Max	11.00	11.30	11.80	11.80	11.48	11.44
Confounding/Aliasing	med	med	low	low	low	low

Figure 1: Design comparison table

Using the example in Figure 1, the team may decide that designs 3, 4, or 5 fit their needs: the number of runs is affordable (30-36) and the power is sufficient (>80%). Design 4 may be chosen because the fraction of the design space (FDS) error is lower than design 3 and does not improve (decreased value) much more in design 5 with the additional points. The variance inflation factor (VIF) is stable between all three designs, and the aliasing is low as well. These design inputs and metrics are detailed below.

NOTE: The matrix can be modified for the specific needs of the test designer and/or for specific design types. This template can be obtained via email at COE@afit.edu.

Typical Design Inputs and Metrics and Their Role in Comparisons

For the design comparison to be meaningful, it is critical for the practitioner to understand the various inputs and how the design metrics are calculated by the DOE software.

- Software Generator/Package
  - Application used to create the designs and allows others to recreate the work
  - Use the same application to compare designs in case differences exist between software packages
- Name/Design type
  - Type of design (fraction, factorial, response surface method (RSM), space filling)

    - Impacts importance of subsequent comparison inputs
- Number of factors
    - Significant to basic design planning and sizing
    - May be a comparison driver if the number of factors is in flux, keeping in mind that removing a factor from the design does not remove its impact on the response
- Number of levels
    - Impacts power as points are divided among factors
    - Implies the number of levels entered by the user into the design software (RSM designs will automatically add more points [center and axial] to be counted in the total number)
    - Can be annotated more specifically for repeatability if needed (e.g. 2x2x3x4 means two 2-level factors, a single 3-level, and a single 4-level factor in a general factorial)
- Model selected
    - Annotated effects intended to be minimally/non-aliased (see below for aliasing) in the final design
    - ME= main effects, 2FI=two factor interactions, Q= quadratic, etc.
- Signal and Noise (ratio)
    - Most DOE software defines this term as a single ratio to calculate effects power; this ratio is not really "selectable" by the test team and creating a design with high power by simply entering a high S/N ratio is a typical mistake
    - May want to avoid this mistake by detailing this ratio by signal (operationally relevant difference to detect) and noise (expected natural variation in the response) and using that ratio in the application
    - Typically the noise may be less understood than the desired signal during the planning phase; in this case, vary the S/N ratio to check its sensitivity with power to see how the design will fair if actual noise is larger than expected
- Alpha (significance level)
    - Set before testing, this determines the confidence level (1-alpha) desired in the reported results
    - Alpha is the false alarm rate for effect significance (typical values are 0.01, 0.05, 0.10, and 0.20)
    - These should not be changed between designs simply to produce a higher test power
- Blocks
    - Blocks represent a group of tests where the differences between blocks represent variables of little interest to the tester
    - Typical DOD blocks are the smallest group of tests conducted together (e.g. days)
    - Blocks consume one degree of freedom per block so including them is critical to good design evaluation
    - The number should reflect the expected number of events per period (e.g. 8 per day, 1 per week)

- Number of center points
    - The presence of center points allows an estimate of pure error (noise) and can be used to check for (not model) curvature in the response
    - While designs can be executed without center points, design sensitivity and analysis goals should be reviewed with center points in mind
    - If all other choices remain the same, the number of centers and replicates can be methodically altered across the matrix to allow a left-to-right flow of increasing test size
- Number of replicate points
    - This definition varies between software packages but generally refers to repeating points in a design to generate an estimate for pure error and build statistical power
    - Models with fewer effects included will require fewer replicate points for the same desired power
    - Increasing replicate points in the comparison allows the practitioner to decide the point at which more replicates are costing more than they are adding in power
- Number runs
    - This tallies the total number of runs for the design
    - This may not be obvious from the number of factors due to the design type (RSM) or the software application generating the points (custom designs)
- Power desired (metric)
    - Power is the ability to correctly identify an effect when it is present and is a function of replicate points, significance level, signal, and noise
    - A rule of thumb is to employ designs with a minimum power of 80% for modeled/desired effects; higher power (beyond the 80% minimum) provides some insurance in case noise is higher than anticipated during actual testing
    - Detailing power for main effects and interactions on different lines in the matrix allows color formatting to be used, but this may require using an average if power varies among effects groups (e.g. main effects).
- Prediction variance (metric)
    - Information regarding the relative predicted variance (with respect to the noise) across the test space is useful in determining if predicted responses will be sufficiently precise for the user's needs
    - A fraction of the design space (FDS) plot may be created by the DOE software to plot the prediction variance across the design space
    - Constant values across the space indicate constant relative prediction variance, and smaller values indicate less relative variance
    - Detailing prediction variance at the 50% and 95% FDS in the matrix allows color formatting to be used and provides some insight to the evaluator regarding how the variance changes across the space
- Variance inflation factor (VIF) (metric)

  - VIF is a diagnostic for detecting multi-collinearity between variables
  - Multi-collinearity means that there is a near linear dependency among the variables Note that polynomial effect terms (X^2, X^3) will expectedly have a high VIF because they are collinear with the main effect (X)
  - Detailing a VIF average and maximum on different lines in the matrix allows color formatting to be used and since VIF can vary across factors, this method provides some insight to the evaluator regarding how the VIF changes across the effects
- Aliasing (metric)
  - Aliasing is the confounding of effects due to a lack of number of points or design orthogonality sufficient to separate effects
  - For classical designs, aliasing is clearly defined by the resolution (III, IV, V, etc). (e.g. resolution III indicates that all main effects are confounded with two factor interactions)
  - For software generated custom designs the resolution is not as clearly defined and is typically reported in a matrix
  - Ideas for assessing the aliasing can be via resolutions (III, IV, V), qualitative (low, med, high), pictorial (pasting pictures of the matrix in the comparison chart), or relatively quantitative (showing aliasing numbers for critical effects)
  - This is difficult to quantify but is important to ensure that desired effects are not confounded

Additional Design Matrix Considerations

- Whole plots
  - Split plot designs (hard to change factors) require whole plot replication to facilitate estimation of whole plot error since the data is statistically separated in the analysis
  - With more than one whole plot factor, it is possible to do an un-replicated factorial or fractional factorial in the whole plots and analyze with a half normal plot
  - Whole plot power comparison can be used to ensure a sufficient number of replicates are being used
- Design Efficiency Diagnostics (D, A, G, V) as seen in NIST (metric)
  - The efficiency diagnostics can be compared as they relate to test objectives
  - D-optimal designs: minimize the generalized variance of the parameter estimators
  - A-optimal designs: minimize the average variance of the parameter estimators
  - G-optimal designs: minimize the maximum variance of the predicted values
  - V-optimal designs: minimize the average variance of the predicted values
- Binary response power predictions
  - Power calculations may be displayed as unrealistically high when the response is binary (pass/fail) because the signal to noise ratio does not directly relate to proportions
  - Accurate S/N ratios and design sizing can be calculated using several methods as mentioned in Ortiz, 2014

Acknowledgements

The original format for this comparison matrix came from Mr. Greg Hutto at Eglin Air Force Base. His long term efforts supporting methodical decision-making are gratefully acknowledged.

References

NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/. 19 Aug. 2014.

Ortiz, Francisco. “Categorical Data in a Designed Experiment Part 2: Sizing with a Binary Response”. Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 2014.