





# Design Evaluation: Statistical Tools for Assessing Your Design Quality


Martin Bezener, PhD  
President & Chief Technology Officer  
[martin@statease.com](mailto:martin@statease.com)

June 2023

1

Making the most of this learning opportunity

 Hide the Control Panel  
 Mute your line



To prevent audio disruptions, all attendees will be muted.

Questions can be posted in the **Question** area. If they are not addressed during the webinar, I will reply via email afterwards.

Questions may also be sent to [stathelp@statease.com](mailto:stathelp@statease.com). Please provide your company name and, if you are using Design-Expert, the serial number (found under Help, About).

**Note:** The slides and a recording of this webinar will be posted on the Webinars page of the Stat-Ease website within a few days.

Design Evaluation

2

## Stat-Ease Training: Sharpen Up Your DOE Skills



### **NEW Summer 2023: ESSENTIALS Series**



**ESSENTIALS 1:**  
Foundations of Factorial  
DOE for Breakthroughs



**ESSENTIALS 2:**  
Strategy of Experiments  
for Accelerating Process  
Development



**ESSENTIALS Series:**  
Response Surface  
Methods for Peak  
Performance

**ESSENTIALS Series:**  
Optimal Tools for  
Formulation  
Development

**Each class: 2 half-days of hands-on learning**

[www.statease.com/training/workshops](http://www.statease.com/training/workshops)

Design Evaluation

3

3

## Agenda



- Setup, motivation, why evaluate
- Power and aliasing for factorial designs
- Fraction of design space for RSM designs
- Practical advice and recap

Design Evaluation

4

4

## Agenda



- **Setup, motivation, why evaluate**
- Power and aliasing for factorial designs
- Fraction of design space for RSM designs
- Practical advice and recap

Design Evaluation

5

5

## Introduction



A typical design of experiments (DOE) workflow **usually** looks something like this:

1. Choose factors and define their ranges
2. Generate an experimental design
3. Evaluate the experimental design
4. Perform the experiment
5. Optimize and draw conclusions
6. Finish or make changes and experiment again

Design Evaluation

6

6

## Introduction



1. Choose factors and define their ranges
2. Generate an experimental design
3. **Evaluate the experimental design**
4. Perform the experiment
5. Optimize and draw conclusions
6. Finish or make changes and experiment again

In practice, step 3 is almost always skipped, **especially** in response surface and mixture experiments!

Design Evaluation

7

7

## Introduction



Performing an experiment without evaluating it beforehand is like....

1. Buying a new car without test driving it first
2. Marrying someone without dating them first
3. Buying something in bulk without trying a small sample

and so on. You get the picture...



Design Evaluation

8

8

## Introduction



- Evaluating a design means examining its properties **before** performing the experiment. Some of the things we look for are
  - high power (can we detect what we are looking for?)
  - good predictive precision (can we get an accurate model?)
  - a design that can provide the information we need
- This can save you from making very costly mistakes
- In my 10+ years of experience, a lot of “poor” DOE results could have been avoided by properly evaluating an experimental design before performing the experiment.
- In this webinar, we will work through two major categories: factorial designs, and RSM/mixture designs.

Design Evaluation

9

9

## Agenda



- Setup, motivation, why evaluate
- **Power and aliasing for factorial designs**
- Fraction of design space for RSM designs
- Practical advice and recap

Design Evaluation

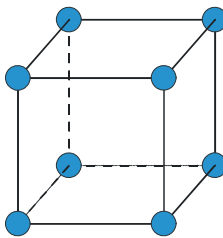
10

10

## Factorial Designs



- A factorial experiment has **k** factors that take one of 2 levels, usually labelled a **low** level and **high** level.
- There is at least one output response that is measured at each combination of the factors in the experiment.
- The goal of factorial experiment is to determine whether any of the **k** factors significantly impact the measured responses between their **low** and **high** levels.



Design Evaluation

11

11

## Factorial Designs



How can we evaluate a factorial design **before** performing the experiment?

- If going from **low** → **high** of a factor significantly impacts the response, are we able to detect this change with high probability?
- Is aliasing an issue if we are not performing a full factorial experiment?
- Is the design fully randomized?
- Is the design possibly too large, or can we better distribute the runs?
- Do we think blocking might be necessary?
- How about center points?

We will go through the **Design Evaluation** tools in our software.

Design Evaluation

12

12

## Example



- This is best illustrated using an **example**. Let's choose a simple one.
- A team is baking cakes and is interested in determining how three factors affect the overall **taste**. Each factor is set to one of two levels:

	<u>Low</u>	<u>High</u>
• Temperature (deg F)	350	400
• Time (min)	25	30
• Milk (cups)	2	2.5



Design Evaluation

13

13

## Example



- Each cake will be independently prepared and baked in the oven by itself. The oven will be cooled down and reheated between each cake.
- After the cakes are prepared, they will be rated (1-9) by a panel of taste testers.
- Some questions:
  - How many cakes do we need to bake?
  - What is the experimental design?
  - What should we check before performing the experiment?
- **First** – let's start building the design in Stat-Ease 360.

Design Evaluation

14

14

## Design Build



### Regular Two-Level Factorial Design

Design for 2 to 21 factors where each factor is set to 2 levels. Useful for estimating main effects and interactions. Fractional factorials can be used for screening many factors to find the significant few. The color coding represents the design resolution: Green (Characterization) = Res V or higher, Yellow (Screening) = Res IV, and Red (Ruggedness testing) = Res III.

Replicates: 1 Blocks: 1 Center points per block: 0 ☐ Show Generators

	2	3	4	5	6	7	8	9	10
4	2 <sup>2</sup>	2 <sup>3-1</sup> <sub>III</sub>							
8		2 <sup>3</sup>	2 <sup>4-1</sup> <sub>IV</sub>	2 <sup>5-2</sup> <sub>III</sub>	2 <sup>6-3</sup> <sub>III</sub>	2 <sup>7-4</sup> <sub>III</sub>			
16			2 <sup>4</sup>	2 <sup>5-1</sup> <sub>V</sub>	2 <sup>6-2</sup> <sub>IV</sub>	2 <sup>7-3</sup> <sub>IV</sub>	2 <sup>8-4</sup> <sub>IV</sub>	2 <sup>9-5</sup> <sub>III</sub>	2 <sup>10-6</sup> <sub>III</sub>

1. Choose the design

A 2<sup>3</sup> factorial design will test all 8 combinations at least once.

2. Enter factor information

### Regular Two-Level Factorial Design

Factors: 3 ☒ Horizontal  
☐ Vertical

	Name	Units	Type	Low	High
A [Numeric]	Temperature	deg F	Numeric	350	400
B [Numeric]	Time	minutes	Numeric	25	30
C [Numeric]	Milk	cups	Numeric	2	2.5

Design Evaluation

15

15

## Power



And then we come to this page...

Responses: 1 (1 to 999)	<input checked="" type="radio"/> Horizontal	<input type="radio"/> Vertical	<input type="button" value="Edit Model..."/>	<input type="checkbox"/> Edit response types
Name	Units	Diff. to detect Delta("Signal")	Est. Std. Dev. Sigma("Noise")	Delta/Sigma (Signal/Noise Ratio)
		2	1	2

How many of you have simply clicked  without paying attention to any of the numbers?

### Design Power

Name	Units	Delta (Signal)	Sigma (Noise)	Signal/Noise	Power for A	Power for B	Power for C
		2	1	2	57.2%	57.2%	57.2%

Design Evaluation

16

16



## Power



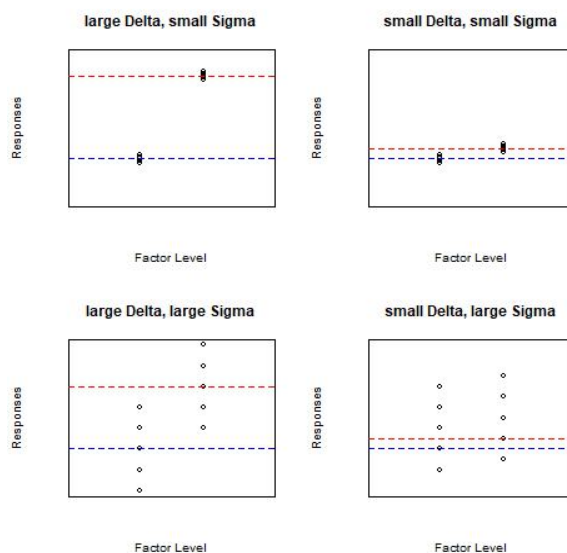
- This screen deals with power, which is a critical statistic in setting up factorial designs.
- Power is about managing expectations – what effects is the experiment capable of detecting?
- Two **key** ingredients that you need to prepare before calculating power:
  - $\Delta$  (delta): the smallest change we are interested in detecting
  - $\sigma$  (sigma): the noise in the process
- Bigger  $\Delta$  are obviously easier to detect, and smaller  $\sigma$  should make detection of active effects easier as well.

Design Evaluation

17

17

## Quadrant of Power



Design Evaluation

18

18

## Power



- If going from the low to high level of a factor truly produces a response difference of  $\Delta$  and the noise in our process is  $\sigma$ , then **power** is the probability of detecting that change.
- We generally want power to be at least 80% for the main effects. Interactions are typically ignored in power calculations.
- In the cake example, we are interested in detecting differences of at least 1 point on a 9-point scale, and we think that noise  $\sigma = 0.5$ , giving a signal-to-noise ratio of  $1/0.5 = 2$ .
- This gives low power but increasing replicates to 2 (16 total runs) bumps the power up to almost 96%.

Design Evaluation

19

19

## Increasing Power



If power comes in low, how can we increase it?

- Add more runs to the design
- Lower the noise in the design - don't just simply choose a lower  $\sigma$ , make an actual effort to lower the noise. Repeated measurements or blocking might be useful here
- Choose a higher alpha cutoff. For example, declare a term significant if  $p < 0.10$  instead of  $p < 0.05$  - **but** this will cause more false alarms.
- Lower expectations and choose a larger  $\Delta$ .

**Remember:** power is about managing expectations. If the power isn't high, you can still perform the experiment and hope for the best. Just don't be shocked if you end up with missed effects.

Design Evaluation

20

20

## Power



### Regular Two-Level Factorial Design

Design for 2 to 21 factors where each factor is set to 2 levels. Useful for estimating main effects and interactions. Fractional factorials can be used for screening many factors to find the significant few. The color coding represents the design resolution: **Green** (Characterization) = Res V or higher, **Yellow** (Screening) = Res IV, and **Red** (Ruggedness testing) = Res III.

Replicates: 2 Blocks: 1 Center points per block: 0 ☐ Show Generators

Responses: 1 (1 to 999) ☒ Horizontal ☐ Vertical  ☐ Edit response types

Name	Units	Diff. to detect Delta("Signal")	Est. Std. Dev. Sigma("Noise")	Delta/Sigma (Signal/Noise Ratio)
Rating	1-9 scale	1	0.5	2

### Design Power

Name	Units	Delta (Signal)	Sigma (Noise)	Signal/Noise	Power for A	Power for B	Power for C
Rating	1-9 scale	1	0.5	2	95.6%	95.6%	95.6%

Design Evaluation

21

21

## Other Issues



- Power is one of the most important statistics to evaluate before performing an experiment, but there are other issues as well.
- A full evaluation can be done after building the design and clicking on the **Evaluation** button.

File Edit View Display Options Design Tools Help

Navigation Pane

- Design (Actual)
- Information
  - Notes
  - Summary
  - Custom Graphs
  - Evaluation**
  - Constraints
- Analysis [+]
  - R1:Rating (Empty)

Std	Run	Factor 1 A:Temperature deg F	Factor 2 B:Time minutes	Factor 3 C:Milk cups	Response 1 Rating 1-9 scale
2	1	350	25	2	
7	2	400	30	2	
14	3	350	30	2.5	
1	4	350	25	2	
11	5	400	25	2.5	
16	6	400	30	2.5	
3	7	400	25	2	

Design Evaluation

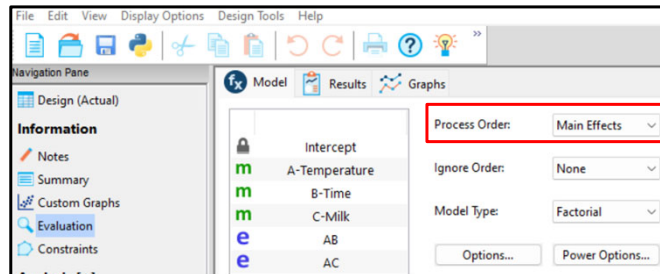
22


22

## Design Evaluation



Select the main effects model:



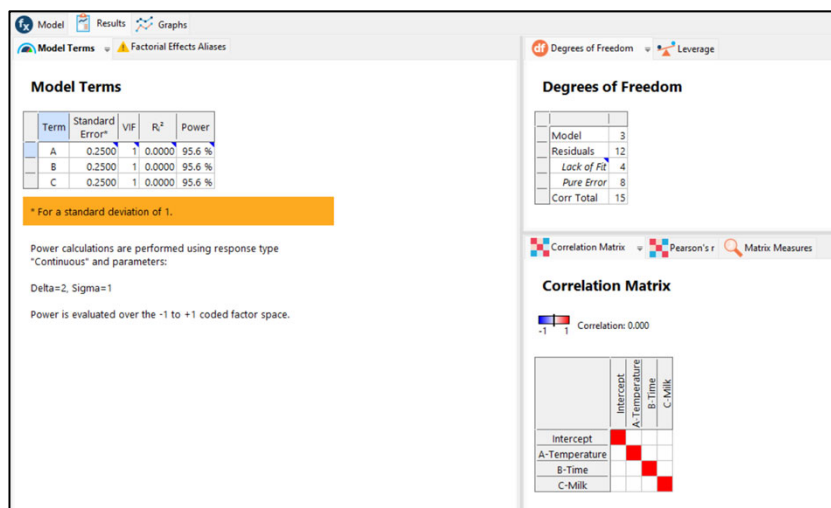
And then click the  Results tab.

Design Evaluation

23

23

## Design Evaluation



Design Evaluation

24

24

## What do these mean?



There's a lot on this screen, and not all of it is super relevant.

- **Power:** already defined, carried over from the design build
- **VIF (variance inflation factor):** how much extra error is introduced due to a factorial design being unbalanced, if for example you do a partial replicate of a design instead of a full replicate.
- **Degrees of freedom:** the amount of information you can extract from the data.
- **Correlation matrix:** shows how 'linked' or correlated estimates of coefficients are

Design Evaluation

25

25

## What about $\sigma$ ?



One of the most common complaints I get is "I have no idea what  $\sigma$  I should use". Common textbook suggestions are

- Use historical or SPC data
- Use an educated guess

Most of the time these suggestions are useless.

**Strategy 1:** Perform an unreplicated factorial design, fit a main effects model, get the estimate of  $\sigma$  from the ANOVA, and add additional runs until power is high enough.

**Strategy 2:** run 4-5 center points, compute the standard deviation of those runs (pure error), multiply by 2 (rough calculation of total error  $\sigma$ ), and then use that  $\sigma$  to build the factorial design around the center points.

Design Evaluation

26

26



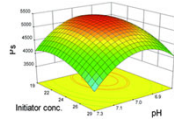
If you have absolutely no idea with regards to  $\Delta$  and/or  $\sigma$ , use these following rules of thumb for spot checking power. They won't be perfect by any means, but they will certainly be better than nothing:

- Plug in  $\Delta = 3$ ,  $\sigma = 1$  if you are just looking to detecting huge, obvious effects
- Plug in  $\Delta = 2$ ,  $\sigma = 1$  if you are looking to detect moderate to large effects
- Plug in  $\Delta = 1.5$ ,  $\sigma = 1$  if you are looking to detect small to moderate effects

Going below a 1.5/1  $\Delta/\sigma$  ratio will result in an explosion of runs with decreasing benefit – the law of diminishing returns kicks in.

- Setup, motivation, why evaluate
- Power and aliasing for factorial designs
- **Fraction of design space for RSM designs**
- Practical advice and recap

## Response Surface Designs



- Factorial designs usually only look at the highs and lows of each factor, possibly at the center point as well. Designs that consider 3 or more levels of each factor are often referred to as **response surface designs**.
- The goal of response surface designs (RSM) is not to detect active effects, but rather fit a precise model (aka response surface) to the data. This response surface can then be used for optimization or making predictions at locations where you didn't collect data.
- Therefore, design evaluation for response surface designs will be different than evaluation for factorial designs.
- **Note:** RSM and mixture design evaluation is very similar, so we will only consider the RSM case here.

Design Evaluation

29

29

## RSM Evaluation



How can we evaluate an RSM design before performing the experiment?

- Is the design fully randomized?
- Is the design capable of mathematically fitting the model that we are interested in?
- Is aliasing an issue? Do we care about any terms that are aliased?
- Will the model have good precision?
- Do we think blocking might be necessary?

We will go through the **Design Evaluation** tools in our software.

Design Evaluation

30

30

## Response Surface Designs



Once again we will work through an example to show how we would evaluate an RSM design before performing the experiment.

A chemical reaction is performed varying the following factors:

	<u>low</u>	<u>high</u>
• time (minutes)	40	50
• temperature (deg C)	80	90
• catalyst (%)	2	3

The response of interest was the yield (%) of the reaction.

Design Evaluation

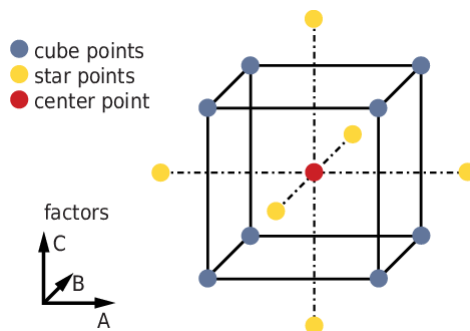
31

31

## Response Surface Designs



The design was set up as a central composite design with 20 runs:



Design Evaluation

32

32



## Response Surface Designs



Std	Block	Run	Factor 1 A:time min.	Factor 2 B:temperat... deg C	Factor 3 C:catalyst %	Response 1 Conversion %
1	Day 1	1	40	80	2	
7	Day 1	2	40	90	3	
10	Day 1	3	45	85	2.5	
12	Day 1	4	45	85	2.5	
4	Day 1	5	50	90	2	
11	Day 1	6	45	85	2.5	
3	Day 1	7	40	90	2	
8	Day 1	8	50	90	3	
6	Day 1	9	50	80	3	
9	Day 1	10	45	85	2.5	
5	Day 1	11	40	80	3	
2	Day 1	12	50	80	2	
14	Day 2	13	53.409	85	2.5	
13	Day 2	14	36.591	85	2.5	
20	Day 2	15	45	85	2.5	
17	Day 2	16	45	85	1.6591	
18	Day 2	17	45	85	3.3409	
16	Day 2	18	45	93.409	2.5	
19	Day 2	19	45	85	2.5	
15	Day 2	20	45	76.591	2.5	

Design Evaluation

33

33

## Response Surface Designs



- **Big Question:** Is this design sufficient? The most important things to check are
  - Aliasing
  - Fraction of design space plot
- Aliasing means you can estimate your desired model “cleanly”. In a response surface experiment, we *usually* want to fit a quadratic model. So we want an experimental design that’s capable of fitting a quadratic model without aliasing/confounding quadratic terms with linear terms.
- Let’s look at an example.

Design Evaluation

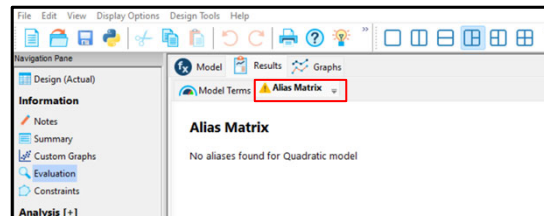
34

34

## Aliasing



Here is the aliasing structure for the chemical reaction RSM design:



Design Evaluation

35

35

## Aliasing



Note that a cubic model would be aliased:

Alias Matrix

There are 6 aliased terms found in the Cubic Model.

Estimated Term	Aliased Terms
Intercept	= Intercept
Day 1	= Day 1
Day 2	= Day 2
A	= A + 2.83 * A <sup>2</sup>
B	= B + 2.83 * B <sup>2</sup>
C	= C + 2.83 * C <sup>2</sup>
AB	= AB
AC	= AC
BC	= BC
A <sup>2</sup>	= A <sup>2</sup>
B <sup>2</sup>	= B <sup>2</sup>
C <sup>2</sup>	= C <sup>2</sup>
ABC	= ABC
A <sup>2</sup> B	= A <sup>2</sup> B + BC <sup>2</sup> - 1.83 * B <sup>3</sup>
A <sup>2</sup> C	= A <sup>2</sup> C + B <sup>2</sup> C - 1.83 * C <sup>3</sup>
AB <sup>2</sup>	= AB <sup>2</sup> + AC <sup>2</sup> - 1.83 * A <sup>3</sup>

Design Evaluation

36

36

## Fraction of Design Space



- We also need to check whether this design is large enough (or too large). Detection of active effects is not the primary goal. Instead, we want a response surface that has high precision, so we use a new tool called **Fraction of Design Space** to evaluate the design.
- **Question:** Have you ever seen the results of a political poll? For example, you may have seen something like “46% of respondents said they prefer candidate **A**, with a +/- 3% margin of error”. What does the margin of error mean?
- The margin of error comes from the fact that we are **not** asking every single voter for their opinion. If you poll a new random sample, you may get 48% or 45% that prefer candidate **A** due to natural variability.

Design Evaluation

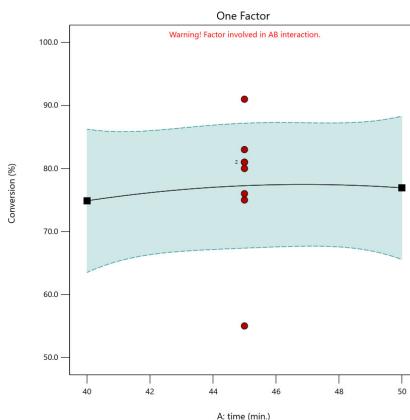
37

37

## Fraction of Design Space



The same issue crops up in RSMs. We fit a model to data and can use it to predict the response at new locations.



Design Evaluation

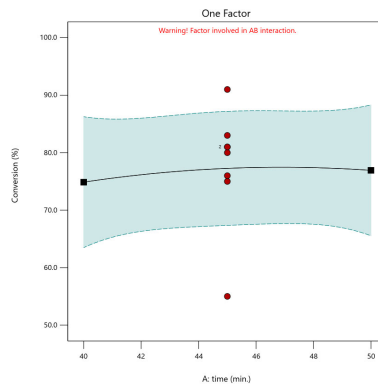
38

38

## Fraction of Design Space



Notice the green intervals. These are called **confidence intervals** and tell us how precise the model predictions are. If we do a run at a factor setting, we should, loosely speaking, expect it to be in or near the confidence interval.



Design Evaluation

39

39

## Fraction of Design Space



**The question is:** can we get the confidence intervals to be narrower? Yes, we can, but that generally requires more data and/or less noise in the process. This is analogous to power in the factorial case.

To compute the precision of a design, we need two ingredients:

- The half-width  $h$
- The noise in the process  $\sigma$

To get the half width, **answer this question:** suppose you optimize my process (e.g. maximize it). What is the +/- uncertainty you'd be okay with around that maximum?

Design Evaluation

40

40

## Fraction of Design Space



- If you maximize a process and the maximum yield comes out to 70%, are you okay with a +/- uncertainty of 30%? How about 10%? 5%? This value is the half width **h** you will need.
- The noise is the same as in the factorial case.
- We generally want the design to produce confidence intervals of +/- **h** in at least 90% of the design space.

**Demo:** We will now evaluate the precision of the chemical reaction design using a half-width of 4% and a  $\sigma$  of 1%.

Design Evaluation

41

41

## Street Smarts



If you have absolutely no idea with regards to **h** and/or  $\sigma$ , use these following rules of thumb for spot checking precision. They won't be perfect by any means, but they will certainly be better than nothing:

- Plug in  $h = 3$ ,  $\sigma = 1$  if you are just looking to protect against massive issues
- Plug in  $h = 2$ ,  $\sigma = 1$  if you are looking for moderately precise models
- Plug in  $h = 1.5$ ,  $\sigma = 1$  if you are looking for very precise models

Just like in the factorial case, going below a 1.5/1  $h/\sigma$  ratio will result in an explosion of runs with decreasing benefit (law of diminishing returns kicks in).

Design Evaluation

42

42

## Agenda



- Setup, motivation, why evaluate
- Power and aliasing for factorial designs
- Fraction of design space for RSM designs
- **Practical advice and recap**

Design Evaluation

43

43

## Practical Advice and Suggestions



Here are some take-home points

- **Always** evaluate a design before performing the experiment
- If you are using a factorial design, aim for 80% power for the main effects.
- If you are using an RSM or mixture design, aim for 95% FDS plot.
- Go through each run in the spreadsheet and perform it “mentally” - make sure it won’t give useless data or cause an explosion. The software doesn’t know about your specific process!
- When in doubt, don’t use your entire run budget in a single pass of the experiment, run a small pilot/calibration experiment and use the information from it to do a second pass of runs.

Design Evaluation

44

44

## Further resources



Stat-Ease offers several free and paid resources that deal with the issue of design evaluation

- Stat-Ease YouTube Channel
- Tutorials / Help System
- Live Hands-On Workshops
- and more! Check out [www.statease.com](http://www.statease.com)

**Final point:** Learning about design evaluation doesn't take long, but it can help in all your future experimentation – the ROI is massive!

45



**Thanks for listening!**

[martin@statease.com](mailto:martin@statease.com)

46