

Designing Experiments for Manufacturing: A DOE Field Guide

OFAT vs. factorial designs. Screening, optimization, and response surface methods. How to design experiments that fit a production schedule, survive real-world constraints, and produce actionable results.

Ref: Montgomery DOE, 9th Ed. · Box, Hunter & Hunter · NIST/SEMATECH · Wu & Hamada

I — Publication

Designing Experiments for Manufacturing: A DOE Field Guide

OFAT vs. factorial designs. Screening, optimization, and response surface methods. How to design experiments that fit a production schedule, survive real-world constraints, and produce actionable results.

Ref: Montgomery DOE, 9th Ed. · Box, Hunter & Hunter · NIST/SEMATECH · Wu & Hamada

Date 2026-03-05

Author Eric Wolters

Reading 20 min

DOE · Factorial

II — Why Experiment

The OFAT Problem

One-factor-at-a-time (OFAT) experimentation is the default approach in most manufacturing operations. Change one variable, hold everything else constant, measure the response. It is intuitive, simple to explain, and wrong in the presence of interactions. If temperature and pressure jointly affect yield in a way that neither does alone, OFAT will never find it. The interaction is invisible because the experiment never creates the conditions that reveal it.

What Factorial Designs Buy You

A factorial design varies all factors simultaneously according to a structured pattern. A 2^3 design (three factors, two levels each) requires 8 runs and estimates all main effects and all two-factor interactions. The OFAT equivalent—varying each factor one at a time with one baseline run—requires 7 runs and estimates only main effects. The factorial costs one additional run and provides three

interaction estimates plus a more efficient estimate of each main effect (because every observation contributes to every estimate).

The cost argument for factorial designs is simple: they extract more information per experimental run. In manufacturing, where each run consumes material, machine time, and operator attention, the efficiency of factorial designs is not a theoretical nicety. It is a direct cost reduction.

III — Screening Designs

Full Factorials: 2^k

A full 2^k factorial tests every combination of k factors at two levels. For $k = 4$, this is 16 runs. For $k = 5$, 32. For $k = 7$, 128. The run count doubles with each additional factor, which becomes impractical above 5 or 6 factors. Full factorials are the right choice when you have few factors ($k \leq 5$), can afford the runs, and want clean estimates of all interactions.

Fractional Factorials and Resolution

A fractional factorial takes a carefully selected subset of the full factorial. A 2^{7-3} design runs 16 of the 128 possible combinations, reducing cost by 87%. The price is aliasing: some effects are confounded with each other and cannot be estimated independently. The **resolution** of the design determines what is confounded with what.

Resolution III: Main effects are aliased with two-factor interactions. You can estimate main effects only if all interactions are negligible. Use for initial screening of many factors (8+) when you expect most factors to be inert.

Resolution IV: Main effects are clear of two-factor interactions, but two-factor interactions are aliased with each other. You can estimate main effects cleanly and detect which interactions are active, but you may not be able to identify which specific interaction is responsible.

Resolution V: Main effects and two-factor interactions are all estimable independently. This is the minimum resolution for optimization work where interactions matter.

Plackett-Burman Designs

Plackett-Burman designs are Resolution III screening designs that exist in run counts that are multiples of 4 (12, 20, 24, etc.). A 12-run Plackett-Burman screens up to 11 factors. They are maximally efficient for screening but provide no interaction information. Use them when you have many potential factors, limited run budget, and the sole objective is to identify the vital few.

IV — Response Surface Methods

CCD and Box-Behnken

After screening identifies the 2-4 critical factors, response surface methods map the relationship between factors and response in detail. The **Central Composite Design (CCD)** augments a factorial with axial (star) points and center points, enabling estimation of quadratic effects. The CCD is the workhorse of process optimization: it detects curvature, estimates the optimum, and requires relatively few runs (a 3-factor CCD needs 20 runs with 6 center points).

The **Box-Behnken design** is an alternative that avoids extreme factor combinations (no corner points). For 3 factors, it requires 15 runs. Box-Behnken is preferred when the corner points are physically impossible or dangerous—for example, when running simultaneously at maximum temperature and maximum pressure would damage the equipment.

Finding the Optimum

The fitted response surface model is a second-order polynomial in the factors. The predicted optimum is found analytically (for unconstrained problems) or numerically (when constraints are active). Contour plots of the fitted surface reveal the shape of the response landscape: flat regions where the response is insensitive to factor changes, steep regions where small changes have large effects, and saddle points where the surface curves in opposite directions along different axes.

Center Points and Blocking

Center points (factor levels at the midpoint of each range) serve three purposes: they test for curvature (a significant center point effect indicates the linear model is inadequate), they provide a pure error estimate for lack-of-fit testing, and they are easy to run because they use nominal settings. Include at least 3-6 center points in any factorial or response surface design.

Blocking partitions the experiment into groups (blocks) that can be run under homogeneous conditions—same day, same material lot, same operator. The block effect is estimated and removed from the analysis, improving the precision of factor effect estimates. Design generators for fractional factorials can be chosen so that blocks are confounded with high-order interactions that are unlikely to be significant.

V — Running the Experiment

Fitting DOE into Production

The textbook says "randomize run order." The production floor says "we cannot change the furnace temperature 16 times in a shift." Both are right. The solution is to acknowledge the constraint explicitly and choose a design that accommodates it.

Split-plot designs handle hard-to-change factors. If furnace temperature takes 2 hours to stabilize, run all combinations of the easy-to-change factors (feed rate, hold time) at each temperature setting before changing temperature. The analysis must account for the split-plot structure; analyzing a split-

plot as a completely randomized design underestimates the error for the hard-to-change factor and inflates its significance.

Randomization When You Cannot Randomize Everything

Full randomization protects against unknown time-trending variables (tool wear, ambient temperature drift, operator fatigue). When full randomization is impractical, randomize within blocks and restrict randomization across blocks. Document what was randomized and what was not. The analysis should match the randomization structure. A designed experiment with unknown restrictions on randomization is an observational study with a structured factor arrangement—still useful, but the inferential strength is reduced.

VI — Analysis

Main Effects, Interactions, and ANOVA

The ANOVA table partitions total variation into components attributable to each factor, each interaction, and residual error. The F-ratio for each term tests whether that effect is larger than expected from noise alone. The p-value gives the probability of observing an effect as large as the estimated one if the true effect is zero. Effects with $p < 0.05$ are conventionally declared significant.

The effect size matters more than the p-value. A factor that is statistically significant but shifts the response by 0.1% of the specification range is not practically important. Report effect estimates in physical units alongside significance tests. A half-normal plot of the absolute effects helps identify the few large effects that dominate the response, especially in unreplicated factorials where formal significance tests are not available.

Residual Diagnostics

Before trusting the ANOVA, check the residuals. Normal probability plot: residuals should fall approximately on a straight line. Residuals vs. fitted values: look for patterns (funnel shape indicates non-constant variance, curvature indicates model inadequacy). Residuals vs. run order: trends indicate time-dependent effects that the randomization did not eliminate. If the residuals show problems, the model is wrong or incomplete, and the ANOVA results are unreliable.

Confirmation Runs

The model predicts optimal factor settings. Run 3-5 confirmation runs at the predicted optimum and compare the observed response to the prediction interval. If the confirmation runs fall within the prediction interval, the model is adequate for its intended purpose. If they do not, the model is missing something—a nonlinear effect, an uncontrolled variable, or an interaction with a factor not included in the experiment. Do not skip confirmation runs. The model is a hypothesis; confirmation tests it.

VII — Common Failures

Too Many Factors at Once

Starting with 12 factors in a Resolution III Plackett-Burman is appropriate for screening. Starting with 12 factors in a Resolution V design (4,096 runs) is not. The sequential approach works: screen with a resolution III or IV design to identify the vital few, then optimize with a response surface design on 2-4 factors. Two small experiments produce better results than one enormous experiment.

Inadequate Replication

An unreplicated 2^4 factorial (16 runs, no replication) has no pure error degrees of freedom. Significance testing requires assuming that higher-order interactions are negligible and using them as the error estimate. This is often reasonable but is an assumption. If center points are included, they provide a pure error estimate. If replication is feasible, even partial replication (replicating a subset of runs) substantially improves the analysis.

Ignoring Blocking Structure

If the experiment was run across two days with 8 runs per day, day is a block. If the analysis ignores the block effect, day-to-day variation inflates the residual error, reducing the power to detect factor effects. Worse, if the factor settings are not balanced across blocks, the block effect biases the factor estimates. Always include the blocking variable in the model, even if you do not care about the block effect itself.

Analyzing Effects Without Checking Residuals

The ANOVA p-values assume normally distributed, constant-variance, independent residuals. If the residuals violate these assumptions, the p-values are unreliable. A transformation (log, square root) often stabilizes the variance and improves normality simultaneously. Check residuals before interpreting the ANOVA table. This takes five minutes and prevents wrong conclusions.

VIII — Related Instruments

INSTRUMENT	SPECIFICATION
Sample Size Calculator	Required n for hypothesis tests and DOE power calculations.
Cpk / Ppk Calculator	Post-experiment capability assessment of the optimized process.
Pareto Chart Generator	Prioritize factors by effect magnitude from screening experiments.

[← All Whitepapers](#)

SVEND is the consulting practice of Eric Wolters – operational excellence, quality management and strategic planning. ILSSI accredited, SCMEP Technical Partnership Provider.

svend.ai/whitepapers/doe-field-guide/ · published 5 March 2026