Screening Ingredients Most Efficiently with Two-Level Design of Experiments (DOE)

A DOE on machine-made bread shows how clever application of statistical methods quickly screens alternative ingredients to see which, if any, impair the desired reaction. In today's extremely competitive world it boils down to the "knead for speed" in making more and more "dough."

Mark J. Anderson

Two-level factorial designs are ideal for identifying the vital few variables that significantly affect your process or product, leading to breakthrough improvements (see sidebar for example). If you want to study 2 or 3 factors, it makes sense to perform all runs in the full two-level design matrix, which consists of all combinations of each factor at its high and low levels. However, if you set your sights to 4 factors and beyond, you may need to cut costly runs by doing only a fraction of all possible combinations.

This article details my attempts to tame my home bread-making machine with fractional two-level design tools. The results at times came out half-baked, but in the end I finally got a rise out of the bread. The exercise provided much food for thought on the statistical repercussions of DOE methods.

First DOE: Sifting Through Flour and Other Ingredients for Making Bread

Modern bread-making machines make it easy to bake your own bread at home. For years I produced machine-made bread using pre-mixes, but recently I decided that it would be more economical (and fun) to buy the ingredients separately and do some baking experiments. I wanted to see if I could save money by using the regular (and cheaper) varieties of flour and/or yeast, as opposed to those that were specifically advertised for bread-making machines. Also, according to a recipe I found printed on the bread-flour package, I could economize by using margarine and water versus butter and milk as fluids. I assumed that any combination of these ingredients would work, and that none of my family members would notice. To be sure, I set up a DOE on the four main ingredients (with cheap versus costly choices coded minus (–) and plus (+), respectively):

A. Liquid: Water (–) or Milk (+)

B. Oil: Margarine (–) or Butter (+)

C. Flour: Regular (–) or Bread (+)

D. Yeast: Regular (–) or Bread (+)

Baking the 16 loaves required for the full factorial would be wasteful, especially if nothing perceptibly changed, so I chose a standard half-fraction requiring only 8 runs (see reference 1 for statistical details on this form of DOE). A design like this works well if nothing comes out significant, or you see only main effects of the test factors, but you lose resolution on interactions of factors (more on this later!). It boils down to a trade-off of experimental runs versus information. In this case I thought none of the factors would be significant so it made sense to choose the lower-resolution half-fraction screening design.

For the first run I added the liquid, oil, flour and yeast at specified levels, as well as salt and sugar (according to the recipe), and set the machine to bake overnight. The resulting bread looked good, but my taste panel (three daughters, one exchange student and my wife) did not like it much (average rating of 4.5 on a scale of 1 to 10). This would not do! I needed to adjust the recipe to get in a more desirable range of taste or face a possible strike by my tasters. I'd endured bad reviews in a previous DOE on baking a pound cake (1). In that case my family responded much better to mixtures with a maximum level of sugar, so I upped this ingredient from 2 <u>tea</u>spoons to 2 <u>table</u>spoons. Then I started over with the DOE. The reviews sweetened up considerably as you can see from the following table of results (shown in standard order, but actually performed in randomized run order).

Table 1.

First Breadmaking DOE: Half-Fraction Two-Level Factorial

Std	A:Liquid	B:Oil	C:Flour	D:Yeast	Taste (avg)	Rise
1a,b	Water	Butter	Regular	Regular	4.7, 5.5	0, 0
2a,b	Milk	Butter	Regular	Bread	6.5, 6.2	1, 1
3a,b	Water	Margarine	Regular	Bread	5.0, 5.2	0, 0
4	Milk	Margarine	Regular	Regular	5.5	1
5	Water	Butter	Bread	Bread	5.8	1
6	Milk	Butter	Bread	Regular	5.0	1
7	Water	Margarine	Bread	Regular	5.3	1
8	Milk	Margarine	Bread	Bread	5.2	1

However, I was shocked to discover that in some cases the bread failed to rise (see picture).

Figure 1.

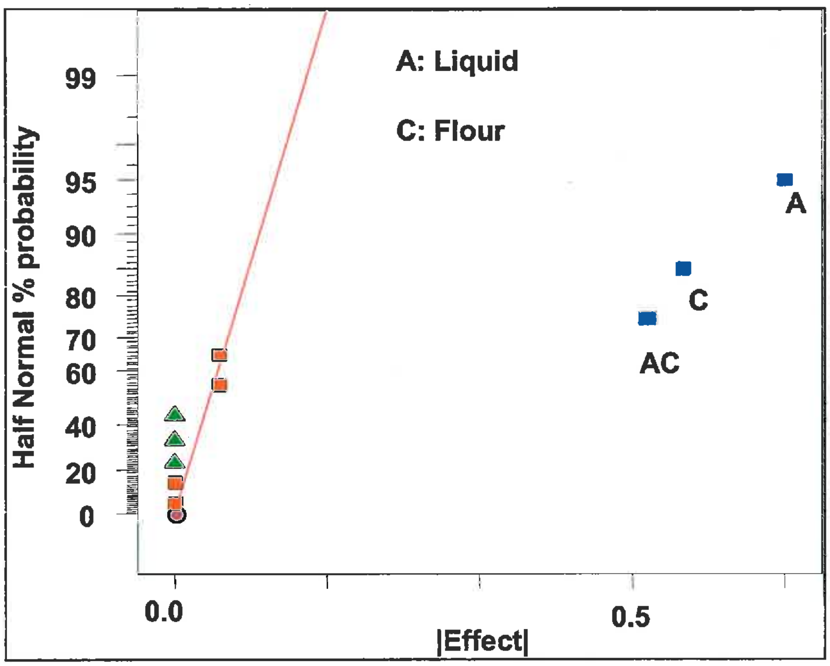
Author Inspects (Half-Baked) Bread

At first I figured that the failure occurred by chance, so I repeated the two runs that failed (standard numbers 1 and 3) and one that didn't (#2). I got the same results, bad and good. Now I knew that the problem could be reproduced.

Which of the tested ingredients, if any, caused the yeast to slack off? To get a numerical analysis, I entered 1's for risen bread ("Yes") and 0's for the bricks ("No"). Then with the aid of statistical software for DOE (3), I produced a statistical plot (see Figure 2), called a "half-normal probability plot," that revealed abnormally large (and highly significant) effects due to the liquid (A) and the flour (C) and the interaction of these two factors (AC). Notice how the other effects (unlabeled squares) and estimates of error (triangles) all fall on a line near the zero-effect level: The y-axis on the half-normal plot is scaled in a special way to make normally distributed effects line up in this fashion. The effects of A, C and interaction AC significantly exceeded what would normally be expected due to random variation.

Figure 2.

Half-Normal Plot of Effects Reveals Interaction



However, things are not as clear-cut as they appear from the plot: Due to the nature of fractional design, interaction AC is *aliased* with BE. This is the price you pay by cutting out half the runs. Let's see what it means to be aliased.

The Pitfalls of Doing Fractional Two-Level Factorial Design

Take a look at Table 2, which lays out the experimental design in coded levels with all interaction columns included. The factor coding is simple: minus (–) for one level versus plus (+) for the other. For numerical factors such as time or temperature, these codes would be assigned to the low and high levels, respectively, but in this case the factors are categorical, so assignment of minuses and pluses is arbitrary (I used cost as the key). The coding for interaction columns is done by multiplying parent terms. For example, the AC column is computed by multiplying the A by the C column.

Table 2.

First Breadmaking DOE: Design Layout in Coded Levels with Interactions Shown

Std	A	B	C	D	AB	*AC*	AD	BC	*BD*	CD	ABC	Rise
1a,b	–	–	–	–	+	+	+	+	+	+	–	0, 0
2a,b	+	–	–	+	–	–	+	+	–	–	+	1, 1
3a,b	–	+	–	+	–	+	–	–	+	–	+	0, 0
4	+	+	–	–	+	–	–	–	–	+	–	1
5	–	–	+	+	+	–	–	–	–	+	+	1
6	+	–	+	–	–	+	–	–	+	–	–	1
7	–	+	+	–	–	–	+	+	–	–	–	1
8	+	+	+	+	+	+	+	+	+	+	+	1

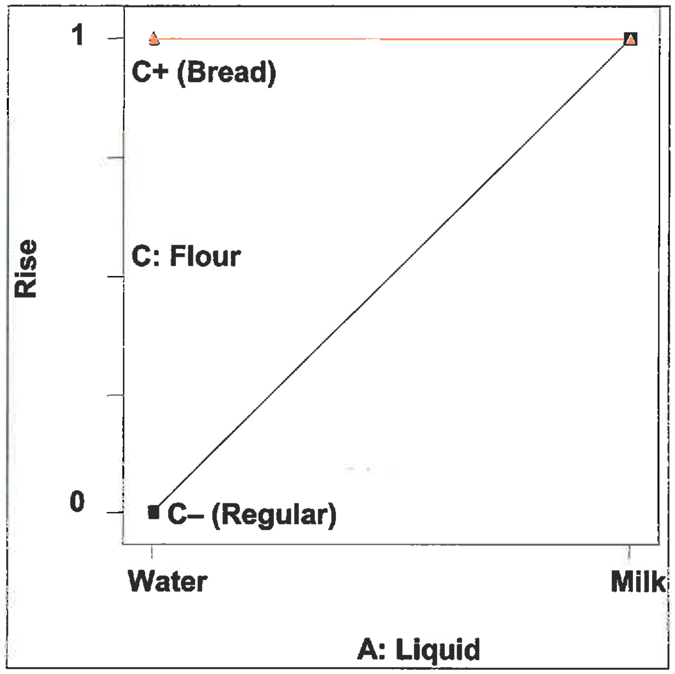
The peculiar thing about this matrix is that A times C equals B times D – these two interactions are therefore statistically "aliased." It's impossible to say which one is really causing the significant effect on bread-making performance because they change back and forth from one level to the other in exactly the same pattern (see italicized columns in the table). Upon closer inspection, you'll notice that all of the two factor interactions are aliased: AB = CD and AD = BC. It also turns out that all main effects get aliased with a three-factor interaction (for example: A = BCD, not shown in the table above, but easily computed by multiplying BxCxD), but it's a generally acceptable practice to ignore interactions of three or more factors, because they're so unlikely. Statisticians characterize this level of aliasing as "resolution IV."

To help you grasp the concept of resolution, think of main effects as 1 factor and add this to the number of interacting factors it will be aliased with. Resolution IV indicates a 1-to-3 (example: A=BCD) and 2-to-2 aliasing (AB=CD, etc.), both of which add to 4 (2+2 and 1+3, respectively). It's possible to create resolution III designs, for example by testing 7 factors in only 8 runs. Then you'd alias main effects with two-factor interactions (1+2=3), which wouldn't be good. If you can afford more runs, choose a design with at least a resolution V, such as 5 factors in 16 runs. Then main effects get aliased only with extremely unlikely interactions of four-factors (1+4=5) and two-factor interactions are confused only with three-factor interactions (2+3=5). However, a resolution IV design often offers a reasonable compromise for screening purposes, which is what I needed for my bread-making experiments.

In this case the only alias of concern was AC=BD, but since neither of the parents (B and D) of the BD interaction came out statistically significant, and these factors (oil and yeast) seemed unlikely to interact, I was tempted to take a leap of faith and place my bets on the AC interaction (Figure 3, shown below, with triangles representing bread flour and squares symbolizing the regular flour).

Figure 3.

Interaction of Liquid (A) with Flour (C) Shows Apparent Problem with Rising



The combination of water and regular flour appeared to create problems with the bread-maker (zero rise at these conditions), but not being a gambling man, I felt it would be best to perform follow-up runs to separate this interaction effect (AC) from its alias BD. To resolve this problem, I made use of a nifty DOE method called a "semifold", which requires adding only half the runs of the original design (4). This technique was developed specifically to improve resolution of two-level fractional factorial designs, such as that used for bread-making, with aliased two-factor interactions (2fi's).

Semifolding the Bread DOE to Resolve Aliased Interactions

Before getting into the details of semifoldover, let's go back a step and talk about full foldover – an established method for enhancing resolution III designs. It's very simple to perform: Just repeat the original experiment with all factors at opposite levels. For example, see Table 3, which shows the layout for a highly-fractionated two-level design on in-line skates (4). Notice that the second block of runs goes opposite the first on all levels.

Table 3.

Full Foldover on Experiment with In-line Skates

Std	Block	A: Pad	B: Bearing	C: Gloves	D: Helmet	E: Wheels	F: Covers	G: Neon	Time (sec.)
1	1	Out	Old	On	Front	Soft	Off	Off	195
2	1	In	Old	On	Back	Hard	Off	On	192
3	1	Out	New	On	Back	Soft	On	On	200
4	1	In	New	On	Front	Hard	On	Off	165
5	1	Out	Old	Off	Front	Hard	On	On	190
6	1	In	Old	Off	Back	Soft	On	Off	195
7	1	Out	New	Off	Back	Hard	Off	Off	166
8	1	In	New	Off	Front	Soft	Off	On	201
9	2	In	New	Off	Back	Hard	On	On	175
10	2	Out	New	Off	Front	Soft	On	Off	211
11	2	In	Old	Off	Front	Hard	Off	Off	202
12	2	Out	Old	Off	Back	Soft	Off	On	205
13	2	In	New	On	Back	Soft	Off	Off	212
14	2	Out	New	On	Front	Hard	Off	On	175
15	2	In	Old	On	Front	Soft	On	On	204
16	2	Out	Old	On	Back	Hard	On	Off	201

Foldovers like this work nicely to improve resolution III designs. However, for resolution IV designs we recommend something a bit different, called a "semifold," which requires that you:

1. Lay out a <u>single-factor</u> foldover from the original design. (Suggestion: choose a factor that's involved in the largest significant two-factor interaction that's aliased with other 2fi(s).)
2. Perform only half of the foldover runs by selecting those where the chosen factor is either at its low level or high level, whichever you believe will generate the most desirable response(s).

Table 4 shows how I applied a semifold to my bread-making DOE. To de-alias the AC interaction, I chose factor C (flour) as the single column I folded over (notice how levels go opposite from the original block of runs shown in Table 1). Then I performed only half the laid-out runs– the ones with regular flour (because I am cheap!).

Table 4.

Semifold on Bread-making Experiment (second block only)

Std	A:Liquid	B:Oil	C:Flour	D:Yeast	Taste	Rise
9	Water	Butter	Bread	Reg		
10	Milk	Butter	Bread	Bread		
11	Water	Marg	Bread	Bread		
12	Milk	Marg	Bread	Reg		
13	Water	Butter	Reg	Bread	5.0	0
14	Milk	Butter	Reg	Reg	6.5	1
15	Water	Marg	Reg	Reg	5.5	0
16	Milk	Marg	Reg	Bread	5.5	1

Now it could be seen that the combination of water and regular flour caused the bread-making to fail (zero rise). As shown in Table 5, this finding is unequivocal because the semifold of four runs de-aliased the interaction of factors A and C from that of B and D. As you can see in the underlined, italicized columns, the patterns no longer match.

Table 5.

Second Bread-making DOE: Design Layout in Coded Levels with Interactions Shown

Std	A	B	C	D	AB	<u>*AC*</u>	AD	BC	<u>*BD*</u>	CD	Rise
13	–	–	–	+	+	+	–	+	–	–	0
14	+	–	–	–	–	–	–	+	+	+	1
15	–	+	–	–	–	+	+	–	–	+	0
16	+	+	–	+	+	–	+	–	+	–	1

Therefore, I concluded that the interaction of factors A and C, depicted in Figure 2, accurately described what affected the bread-making process. I must avoid the combination of regular flour and water. That's not a problem, because with a family like mine, there's always milk in the refrigerator, so I just use it instead of water and the bread always rises. I use margarine and regular yeast with the regular flour to hold keep ingredient costs to a minimum. Unfortunately, due to the unexpected failures in getting my bread to rise, I lost sight of my original objective: Improve taste. The statistical analysis does show a tendency to prefer the same conditions that resulted in risen breads. However, some of my children actually rated the failed breads higher, which created ambiguity in the findings. They must like the gooey mouth-feel of dough (Yuk!). My follow-up studies will likely involve not only what ingredients to use, but also how much of each. Stay tuned for further revelations on the mysteries of making tasty product from bread-making machines!

Lessons Learned

Let's go back over this series of bread-making experiments and see what can be learned from the experience. Despite all the complications in this case, I would not hesitate to choose a resolution IV design for screening purposes, because they allow you to study many factors in few runs. For example, you can study up to 8 factors in only 16 runs and still get resolution IV. Good DOE software (3) lays out a full array of designs of varying resolution for many more factors (if you can handle them) and many more runs (if you can afford them). What you do as a result of running a resolution IV screening design depends on which, if any, effects come out significant. Here's a general strategy for follow-up:

- Scenario 1 - Nothing significant: Look for other factors that affect your response(s).
- Scenario 2 - Only main effects significant: Change these factors to their best levels.
- Scenario 3 -Two-factor interaction(s) significant: De-alias by performing a semifold.

By following this strategy you will increase your odds of uncovering breakthrough main effects and interactions at a relatively minimal cost in experimental runs. This is an ideal situation - akin to baking your bread and eating it too.

References

(1) Anderson, M.; Whitcomb, P.; *DOE Simplified, Practical Tools for Experimentation*, **2000**, Productivity, Inc., Portland, Oregon.

(2) Anderson, M.; Whitcomb, P.; "Mixing it Up with Computer-Aided Design," *Today's Chemist,* November, **1997**, pp. 34-38.

(3) Design-Expert® software, Version 6, **2000**, Stat-Ease, Inc., Minneapolis (www.statease.com).

(4) Anderson, M.; Whitcomb, P.; *How To Save Runs, Yet Reveal Breakthrough Interactions, By Doing Only A Semifoldover On Medium-Resolution Screening Designs*, **2001**. 55th Annual Quality Congress of the American Society of Quality (Milwaukee, WI.).

Acknowledgments

I wish to thank Patrick J. Whitcomb, my partner at Stat-Ease, for inspiring me to experiment with the DOE methods detailed in this article.

Mark J. Anderson *is a principal at Stat-Ease, Inc., Minneapolis. He earned B.S.Ch.E. and M.B.A. degrees at the University of Minnesota. His e-mail address is Mark@StatEase.com.*

Sidebar: Knead for Speed in Making Herbicide *(Extracted from a Case Study by David Long, Aventis CropScience)*

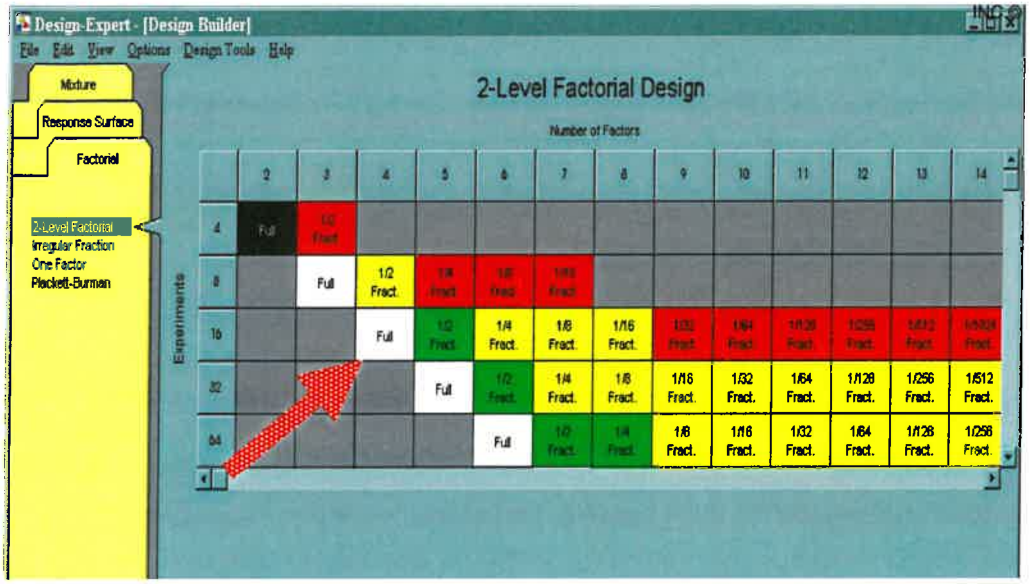
Aventis CropScience (Research Triangle Park, NC) recently introduced an herbicide formulated specifically for the Brazilian corn and sugar cane markets. However, after successful pilot tests they encountered problems in the field due to clogging in the applicators. On of their principal scientists, David A. Long, set up a two-level factorial design to investigate how various factors affected the dispersion of the solid herbicidal material into the aqueous medium. To set up and analyze the experiment, he made use of a DOE software package called Design-Expert® (Stat-Ease, Inc., Minneapolis, MN, 1.612.378.9449, www.statease.com).

The Aventis herbicide is prepared in a process that's similar to making pasta. A powdered form of the herbicide is first mixed with a small amount of water. Then it's kneaded to a dough-like substance and extruded through fine holes. The herbicide is then dried on a vibrating fluidized bed dryer and packaged as small granules measuring 0.8 mm in diameter by 2 to 3 mm in length. David and his project team selected four factors they thought might cause the herbicide to clog in the Brazilian field applicators:

A. Dispersant level
B. Particle size before extrusion
C. Amount of water
D. Extrusion rate.

With the aid of Design-Expert, David set up a design with two levels of each factor and laid out a randomized run plan for their pilot-scale equipment. The screen shot shows Design-Expert's two-level factorial design builder. It exhibits color-coding, similar to a stoplight, that clue users in on safe (green) versus cautious (yellow) versus risky (red) design options. Clicking a box initiates a sequence of steps leading to a "recipe" sheet in randomized run order. The arrow points to the full design on 4 factors (requires 16 runs) – a safe choice for the type of problem faced at Aventis.

SCREENSHOT COURTESY STAT-EASE, INC.



Statistical analysis from the pilot trials revealed a number of significant effects on dispersion. Armed with this information, David and his team set up confirmation runs on a larger facility at what they thought would be ideal extrusion settings. However, they encountered unacceptably high temperatures. Perhaps due to friction caused by particles passing through the extrusion holes. The team of scientists then looked for what differed from pilot plant to full-scale equipment and pinpointed the kneading operation as the likely culprit. After increasing the kneader speed and residence time, the temperature problems disappeared. Aventis then proceeded with their production run and shipped the herbicide to Brazil. It was applied successfully, with no further complaints about clogging in the applicators.

In this case the factorial DOE approach proved to be absolutely essential for determining which factors influenced critical product properties, particularly given the time pressures (the herbicide had to be applied during the Brazilian growing season). Traditional one-factor-at-a-time (OFAT) experiments would not only have taken too long, they would never have revealed powerful interactions of factors, which often prove to be the key to success. Aventis might've lost millions of dollars in foregone sales and disposal costs for unusable product. They would have also lost valuable goodwill and incurred more expense developing a replacement product. David says in conclusion: "We certainly learned more about our product using DOE and Design-Expert than we would have using other methods."